4Humanities "WhatEvery1Says" (WE1S) Project Proposal Narrative & Budget Narrative

(Abreviated Version)

(June 22, 2017)

Table of Contents

Proposal Narrative

I. Description of Project

a. Overview

Based at University of California, Santa Barbara (UCSB), with core collaborators at California State University, Northridge (CSUN) and University of Miami (UM), the "WhatEvery1Says" project (WE1S) uses digital humanities (DH) methods to study public discourse about the humanities at large data scales. The project concentrates on, but is not limited to, journalistic articles available in natively or OCR'd digital textual form beginning circa 1981. Currently in the midst of a small-scale pilot project for this purpose, WE1S proposes to use Mellon funding on a timeline of three years (beginning October 1, 2017) to expand significantly the scope and diversity of its sampled materials; to increase the range, nuance, and trustworthiness of its analytical methods; and to make its technical research environment agile enough to support rapid, flexible exploration of new materials and research questions. WE1S's parent initiative is 4Humanities.org, which PI Alan Liu started in 2010 with international collaborators to use digital technologies for humanities advocacy.

b. Humanities Context

WE1S contributes to recent research responding to the perceived long-term decline of the humanities, including after the most recent “crisis” period touched off by the Great Recession of the late 2000s and early 2010s. Such research has been broad and vigorous. For example,

- Scholars such as Jonathan Bate, Eleonora Belfiore and Anna Upchurch, Rens Bod, Peter Brooks, Geoffrey Harpham, Gordon Hutner and Feisal G. Mohamed, Martha Nussbaum, Helen Small, and Sidonie Ann Smith have written books on the value and history of the humanities *(see section III, Works Cited)*.
- Other scholars in the Society for the History of the Humanities have started the *History of Humanities* journal to publish new historical and comparative research on the humanities.
- The innovative Humanities & Liberal Arts Assessment (HULA) project has studied and assessed the “implicit internal logics of humanistic craft” in order to surface the methods and values of the humanities. (Especially akin to WE1S's focus on discourse about the humanities is the HULA report by Danielle Allen, et al., titled "Humanities Craftsmanship," which studies the characteristics of 30 years of humanities grant applications awarded funding by the Illinois Humanities Council.)
- Major scholarly associations and foundations for the humanities have issued reports, white papers, and policy recommendations (e.g., the American Association of University’s *Reinvigorating the Humanities* [Mathae and Birzer, 2004] and the American

Academy of Arts and Science Commission on the Humanities and Social Science's *The Heart of the Matter* [2013]).[1]

- The American Academy of Arts & Science has created Humanities Indicators and Academy Data Forum to gather significant statistics.
- Meanwhile, "public humanities" initiatives of many varieties along with humanities advocacy initiatives (ranging from nationally organized coalitions such as the National Humanities Alliance to grassroots initiatives such as 4Humanities.org, the parent initiative of the WhatEvery1Says project) have been active in communicating the value of the humanities to the public and its representatives in government and the media.

WE1S adds uniquely to this broader field of research and advocacy by using digital humanities methods--mainly topic modeling *(see section I.e, Research Methods)*--to analyze representations of the humanities in large numbers of public materials, especially journalistic media. If the Humanities Indicators project provides statistical research on the state of the humanities, WE1S provides the other half of the picture: discourse research on how the humanities are articulated in public and at crossover points between the public and the academy.

Specifically, WE1S explores the following research hypotheses, which--depending on results--may lead to iterative, new, or alternative hypotheses:

- That newspaper articles and other documents containing the literal phrases "humanities", "liberal arts", and "the arts" are likely places to look for focused discussion of the humanities (e.g., articles on the "humanities crisis") *and* socially broad discussion of the humanities (e.g., articles on the humanities as part of personal life and general culture);
- That the crossing point between such focused and broader views can help us understand the "architecture" of the "complex idea" of the humanities (to use Peter de Bolla's vocabulary in his *The Architecture of Concepts*, which studies discourse on the analogously complex idea of "human rights");
- That there is a canon of themes, narratives, examples, metaphors, and evidence types used by journalists, educators, politicians, parents, students, and others to weigh public or personal decisions about the humanities;
- That there may be other important themes, narratives, examples, metaphors, and evidence types whose role in public discourse on the humanities is unrecognized or underweighted;
- And that there are differences in the way the humanities are discussed across different media sources, nations, and time; as well as by, or in relation to, different racial, ethnic, gender, immigrant, or age groups.

[1] Other major reports on the humanities are listed under "Resources" on the site of the Commission on the Humanities and Social Science.

c. Digital Humanities Context

As a digital humanities project, WE1S also contributes to the evolving context and methods of the digital humanities field in three ways:

- WE1S takes its place in the evolving branch of the digital humanities called "cultural analytics," which brings into convergence “distant reading,” text analysis, topic modeling, and other data-analytic methods to study sociocultural, historical, and aesthetic phenomena at collectively significant scales. Symptomatic is the advent of the journal *Cultural Analytics*. Individual examples of cultural interpretation based on large-scale analyses of historical textual corpora include studies by Ryan Cordell, Andrew Goldstone and Ted Underwood, Ryan Heuser, Mark Algee-Hewitt, Matthew Jockers, Andrew Piper, Benjamin Schmidt, Richard Jean So and Hoyt Long, and Matthew Wilkens. Similar studies have been conducted based on audio-visual corpora by Tanya Clement; and Lev Manovich, Jeremy Douglass, and William Huber.[2] In addition, WE1S is similar to "new media studies" data-analysis projects in focusing on contemporary media materials in an unclosed, evolving document set. Though it does not study social media, its interest in recent and ongoing journalistic media may be analogized to projects like R-Shief that analyze and visualize Twitter.
- Technologically, WE1S contributes to the development of integrated frameworks for data-analysis workflow by creating an adaptable data workflow system that draws on the principles of more complex digital humanities and scientific workflow systems but streamlines them (and translates the idea of "data provenance" in scientific workflow into that of "document" provenance). This creates a data-analysis workflow system that is more practically usable and intellectually graspable for a larger number of digital humanities scholars. The main examples of high-powered but complex data workflow systems in the digital humanities at present are the SEASR / Meandre workflows in the HathiTrust Research Center, which allow scholars with advanced technical knowledge to work with large numbers of texts in ordered sequences of data preparation and analysis. These tools have not been widely adopted by the digital humanities community, in part because they are difficult to implement and use. The main examples of scientific data workflow systems, which are even more powerful and complex, are Apache Taverna, Kepler, and Wings. These systems have also failed to find users in the digital humanities community for similar reasons. By contrast, WE1S's combined Workflow Management System and Virtual Workspace System (described below) are similar in their usability to such ready-to-go online or installable digital humanities systems as Voyant Tools and DH Box, though unlike these (which are like toolboxes), WE1S is oriented toward allowing researchers to operate structured workflows that chain together tools in series to achieve specific end-goals (e.g., conducting the whole sequence of data ingest, data cleaning, topic modeling, visualization, and ancillary processes that create a topic model of a corpus and present it for interpretative exploration). The closest comparison to the WE1S data workflow environment at present is Lexos, developed by the NEH-funded Lexomics project, which provides an online "integrated lexomic workflow" for text ingest,

[2] Cordell's research into historical American newspapers is especially apropos. WE1S has consulted with Cordell and hopes to include him as well as several other experts mentioned in this proposal section on its advisory board.

cleaning, analysis, and visualization optimized for the individual researcher or small-team digital humanities project.[3] Like Lexos, WE1S relies on technologies that are accessible and familiar to digital humanists (e.g., Web-based markup and scripting and the non-compiled programming language Python). Scott Kleinman, one of WE1S's co-PIs, has been co-PI of the Lexomics project and the developer for Lexos. WE1S anticipates synergies between the two projects--for example, using Lexos to generate visualizations and cluster-analyses of topic models that can assist in interpreting topic models.

- Additionally, WE1S contributes to the development of open, shareable, and reproducible methods in the digital humanities. One of the reasons that the scientific data workflow systems cited above are so powerful (more so than the complex digital humanities workflow systems also cited) is that they are based on open metadata standards for describing data materials and their transformations (such as the W3C PROV protocol for provenance) and also shareable ways of passing such metadata to other systems (such as the JSON format for "serializing" data).[4] The result is that data workflows in the sciences are reproducible--i.e., documented in computationally-tractable ways that allow them to be repeated (and iterated or evolved). Because WE1S's Workflow Management System and Virtual Workspace System not only implement workflows but do so in open, annotated ways (creating provenance "manifests" using JSON and operating on them using open-science Jupyter notebooks[5]), they introduce to the digital humanities the kind of workflow systems based on metadata standards that the *in silico* or data-intensive sciences have advanced under the rubrics of “open science” and “open lab.”[6] This is important to advance the state of scholarship in the digital humanities, where the conventions for publishing not just the conclusions of a data-analysis project but also the underlying data and workflows are now beginning to emerge. For example, the *Cultural Analytics* journal is pioneering for the digital humanities a publication platform and policy that require authors to deposit for open access the data and processing scripts underlying their research articles (where the intellectual property status or size of a dataset permits).[7] WE1S's development of methods for declaring, annotating, and

[3] The Lexomics project was initially formed with the aid of a Mellon Foundation grant to Wheaton College, Massachusetts, to foster interdisciplinary connections in its curriculum and to support student-faculty collaborative research during summer 2007. Its summer research model has continued every year since. The Lexomics group now includes participants from multiple institutions, and its Lexos tool is widely used for teaching and research.

[4] The JSON serialization format organizes information into keyword-value pairs for describing a resource or process. Such files are generally readable by humans, but can also be parsed by computers.

[5] Jupyter notebooks (previously known as "iPython notebooks") are documents stored in the JSON format that can not only narrate data processing steps but run actual code in step-by-step modules.

[6] Recent research on open, shareable, and reproducible data workflows in the sciences includes articles by Daniel Garijo and Yolanda Gil. There has been some early research on digital humanities workflow--e.g., James Clawson, "Who’s Afraid of Topic Modeling? Proposing a Collaborative Workflow"; and Smiljana Antonijievic Ubois and Ellysa Stern Cahoy, "Supporting Humanists’ Digital Workflow" [see Rockwell]. A recent article on the reproduction and reuse of data analysis in the digital humanities is Sarah Allison's “Other People’s Data: Humanities Edition”.

[7] See *Cultural Analytics*'s "Data Sharing Policy," http://culturalanalytics.org/about/about-ca/.

sharing workflows--complete with provenance information and sequences of actions and tools--will add to such emerging scholarly protocols. Such conventions will strengthen the credibility and impact of digital humanities scholarship by allowing data and methods to be examined, tested, and adapted for use by others. In addition, opening the "black box" in which digital humanities studies have often hidden their methods of gathering, cleaning or pre-processing, analyzing, and reaching conclusions about their materials is important in helping to disseminate digital-humanities methods to beginners, scholars in other fields, and tenure promotion committees.

d. Expected Audiences and Outcomes

WE1S aims for its research and methods to serve three overlapping audiences in the following ways:

i. For the public, WE1S will provide research-based examples and analyses of themes, narratives, metaphors, evidence, and value statements about the humanities, together with links to readings in the original journalistic material. WE1S's research will thus complement that of the American Academy of Arts & Sciences' Humanities Indicators project, which provides data and statistics on the humanities.[8] In addition, WE1S will create resources and recommendations to help guide discussion about the humanities by journalists, politicians, business people, university administrators, parents, and students.

ii. For humanities scholars and administrators, WE1S will provide articles, white papers, open metadata, interpreted results, and research workflows and tools representing its project. These can be used for study in such research areas as: university studies; the idea and value of the humanities; the history of the humanities; and "global" or comparative humanities. More broadly, the project will provide methods and tools for humanities researchers investigating the role of complex ideas in society.

iii. For digital humanities scholars, WE1S will contribute methods and tools (to be used either "as is" or in adapted form) for integrated, open, shareable, and reproducible data analysis and interpretation *(as explained in section I.c above)*.

e. Research Methods

WE1S's research starts with identifying and harvesting for analysis documents from journalistic sources (and in the future other sources in the public sphere; *see section II.b.1*) that include the phrases "humanities," "liberal arts," and (in the United Kingdom and Commonwealth nations, "the arts").[9] For example, WE1S's pilot project *(see section II.a)* has gathered data on about 36,000 articles related to the humanities from a small number of high-value, English-language journalistic sources after c. 1981 (i.e., after the advent of fully digitized newspaper source

[8] WE1S has consulted on its plans with Robert B. Townsend, Director of the American Academy of Arts & Science's Washington D.C. office, and hopes to include him on the project advisory board.

[9] WE1S collects articles using the phrase "the arts" in the United Kingdom and the Commonwealth for reasons documented in its study, "How Public Media in the US and UK Compare in Their Terminology For the Humanities." Besides searching on "humanities," "liberal arts," and "the arts," WE1S will experiment with other search patterns and methods in the future *(see sections II.b.2 and II.e.2)*.

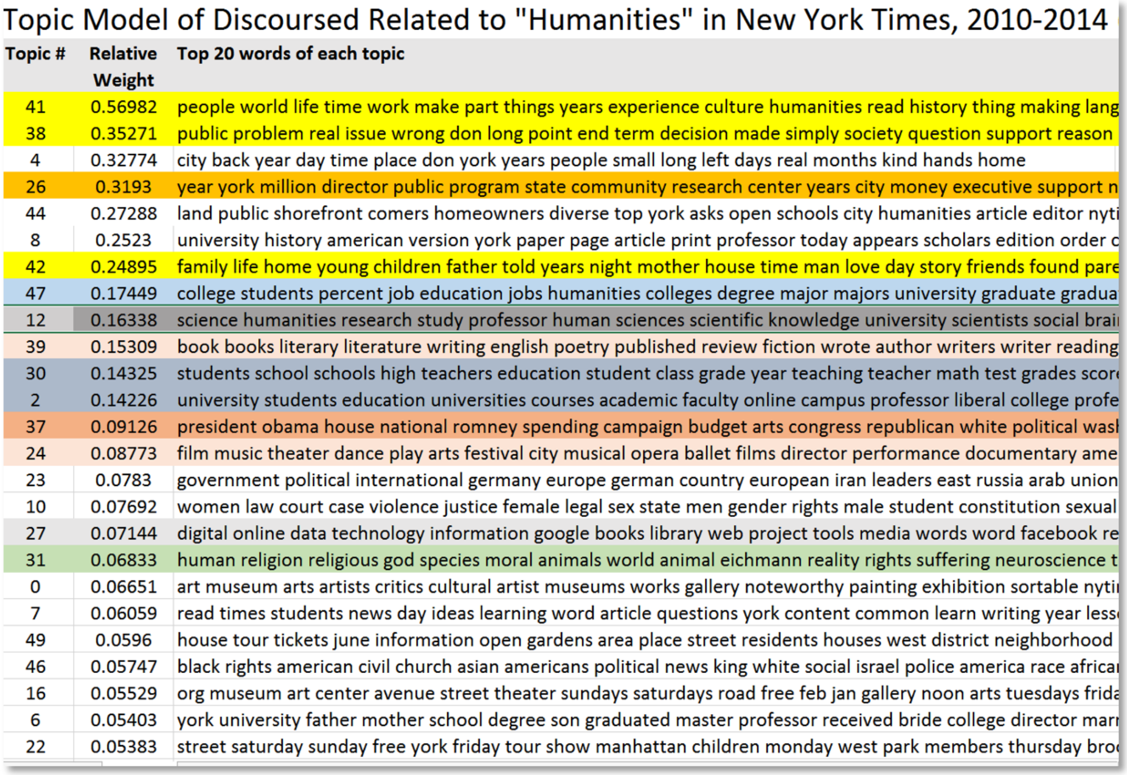
material).[10] Text is "scraped" in plain-text form either directly from a publication's API (application program interface) or from databases (through manual searching and downloading as constrained by licensing restrictions on algorithmic harvesting, followed by automatic scraping). These plain texts are "cleaned," undergo other pre-processing, and are then converted into analytical data for machine learning processes such as topic modeling. "Analytical data" means that, in accordance with non-consumptive use practices, the texts of original articles are not stored. Instead, each article is stored only as an alphabetized "bag of words" file before becoming available to project workflows. Original articles thus cannot be reconstructed from these files.[11] However, metadata about the original documents (e.g., citations and, where possible, links to the original locations of articles in their proprietary or other locations) is stored. Additional proposed methods such as word embedding *(see below in this section)* will require storing other non-consumptive, non-reconstructable document data in the same fashion *(see also under section II.g, Intellectual Property)*.

To allow for null hypothesis testing, WE1S also gathers from its sources analytical data for a smaller "random" corpus of articles. A random rather than "control" sample is used for this purpose because in public discourse there are no natural boundaries between what does and does not count as related to the humanities. For example, the humanities can appear in both precise and general contexts: as a focal topic, as part of arts and culture, in particular forms (e.g., literature), as part of social and ethical concerns, or as part of the biographies or obituaries of individuals. Indeed, it may be that one distinction of the humanities is precisely their capacity to intersect along multiple pathways between tightly focused and general themes. There is thus no pre-definable "control corpus" of public discussion on the humanities that can serve as the "ground truth" for WE1S's research (i.e., a control sample supervised by human readers able to determine intuitively and definitively what constitutes discourse on the humanities). WE1S's random test corpus is relatively small (for the current pilot project approximately 2,000 articles drawn from *The New York Times*, *Washington Post*, and *The Wall Street Journal)*. But it represents a statistically meaningful, year-by-year proportional subset of the project's larger corpus. WE1S will use the test corpus to provide an initial sense of the boundaries of public discourse about the humanities. Applying methods of statistical text classification to compare its main corpus to its random corpus, WE1S may be able to determine algorithmically what differentiates media discourse on the humanities from, for example, such discourse on the sciences, business, or politics. (For more on text classification methods, *see below in this section*).

[10] The pilot project gathered material from six U.S. sources, including major newspapers such as the *New York Times*, *Washington Post*, *Wall Street Journal*, and *Los Angeles Times*; and one major paper each in the United Kingdom and Canada).

[11] "Bags of words" are representations of documents in the form of frequency counts of words and other extracted or derivative data that are "non-consumptive" representations because they do not allow for reading the original documents. As defined, for example, by the HathiTrust (in compliance with fair use rulings bearing on the use of copyrighted materials for machine learning), "Non-consumptive analytics includes such computational tasks as text extraction, textual analysis and information extraction, linguistic analysis, automated translation, image analysis, file manipulation, OCR correction, and indexing and search" ("Non-Consumptive Use Research Policy").

The main computational method that WE1S applies to analyze its gathered materials is topic modeling (specifically, Latent Dirichlet Allocation [LDA] topic modeling as implemented in the standard MALLET toolkit [Machine Learning for Language Toolkit]). A leading method of machine-learning analysis, topic modeling discovers through statistical means the existence, relative weight, and distribution of “topics” across documents (where topics are represented as a probability model of correlated words often indicative of what a human might conceive as "themes"). Widespread adoption and discussion of the method in the digital humanities and such other fields as the digital social sciences have demonstrated its usefulness. Experimental topic models WE1S has produced in its pilot-project by analyzing thousands of newspaper articles have thus already identified various public topics associated with the humanities and differentiated their relative weight (see Figure 1 for a partial view of a topic model, rendered in spreadsheet form, of five years of *New York Times* articles related to the humanities). Topic modeling can be particularly important for discovering areas of public discourse related to the humanities that are not colored by preconceived theses or expectations (e.g., about the "crisis" of the humanities). For example, the topics labeled #23 and #10 in Figure 1, whose frequent words include, respectively, *"government political international germany europe german country european iran leaders, east russia arab union"* and *"women law court case violence justice female legal sex state men gender rights male student constitution sexual,"* may not at first glance seem as predictably related to the humanities as other topics filled with words on education, books, theaters, or museums. Such topics mark out research sweet spots where the



Topic Model of Discoursed Related to "Humanities" in New York Times, 2010-2014

Topic #	Relative Weight	Top 20 words of each topic
41	0.56982	people world life time work make part things years experience culture humanities read history thing making lang
38	0.35271	public problem real issue wrong don long point end term decision made simply society question support reason
4	0.32774	city back year day time place don york years people small long left days real months kind hands home
26	0.3193	year york million director public program state community research center years city money executive support n
44	0.27288	land public shorefront comers homeowners diverse top york asks open schools city humanities article editor nyti
8	0.2523	university history american version york paper page article print professor today appears scholars edition order c
42	0.24895	family life home young children father told years night mother house time man love day story friends found pare
47	0.17449	college students percent job education jobs humanities colleges degree major majors university graduate gradua
12	0.16338	science humanities research study professor human sciences scientific knowledge university scientists social brai
39	0.15309	book books literary literature writing english poetry published review fiction wrote author writers writer reading
30	0.14325	students school schools high teachers education student class grade year teaching teacher math test grades score
2	0.14226	university students education universities courses academic faculty online campus professor liberal college profe
37	0.09126	president obama house national romney spending campaign budget arts congress republican white political wash
24	0.08773	film music theater dance play arts festival city musical opera ballet films director performance documentary ame
23	0.0783	government political international germany europe german country european iran leaders east russia arab union
10	0.07692	women law court case violence justice female legal sex state men gender rights male student constitution sexual
27	0.07144	digital online data technology information google books library web project tools media words word facebook re
31	0.06833	human religion religious god species moral animals world animal eichmann reality rights suffering neuroscience t
0	0.06651	art museum arts artists critics cultural artist museums works gallery noteworthy painting exhibition sortable nyti
7	0.06059	read times students news day ideas learning word article questions york content common learn writing year less
49	0.0596	house tour tickets june information open gardens area place street residents houses west district neighborhood
46	0.05747	black rights american civil church asian americans political news king white social israel police america race africa
16	0.05529	org museum art center avenue street theater sundays saturdays road free feb jan gallery noon arts tuesdays frida
6	0.05403	york university father mother school degree son graduated master professor received bride college director marr
22	0.05383	street saturday sunday free york friday tour show manhattan children monday west park members thursday bro

Figure 1: Partial view of topic model of articles mentioning "humanities" in The New York Times during 2010-2014, ranked in descending order or relative weight. Topics are identified by the most frequent words associated with them. Color coding has been added to distinguish different kinds of topics--for example, general-life topics, higher-education topics, cultural arts topics, and political topics.

topic model might send human interpreters back to access and read some of the original articles contributing to that topic. (For effective introductions to topic modeling written for scholars in the humanities and social sciences, respectively, see Underwood, and Mohr & Bogdanov. For an introduction intended for a general scientific audience by one of its inventors see Blei.)

WE1S interprets topic models by following an interpretation protocol (a repeatable sequence of human reading/interpreting activities syncopated with iterative machine-learning steps) that is currently under evolution *(see under section I.f, Technical Methods)*. The idea is that the interpretation of machine learning results should follow a declarable set of procedures and steps that allow others to understand (and iterate or improve) how a project makes observations and draws conclusions from topic models. Facilitating the interpretive exploration of topic models is WE1S's use of the dfr-browser topic-model visualization interface developed by Andrew Goldstone, which was chosen as optimal after WE1S conducted a comparative study of 14 topic-model interfaces.[12] By comparison with spreadsheet or other tabular, static representations of topic models (of the sort seen in Figure 1), dfr-browser (as seen in Figure 2) is a relatively intuitive and dynamic interface for observing a topic model through different perspectives, including views of the overall set of topics, ranked frequent words in topics,

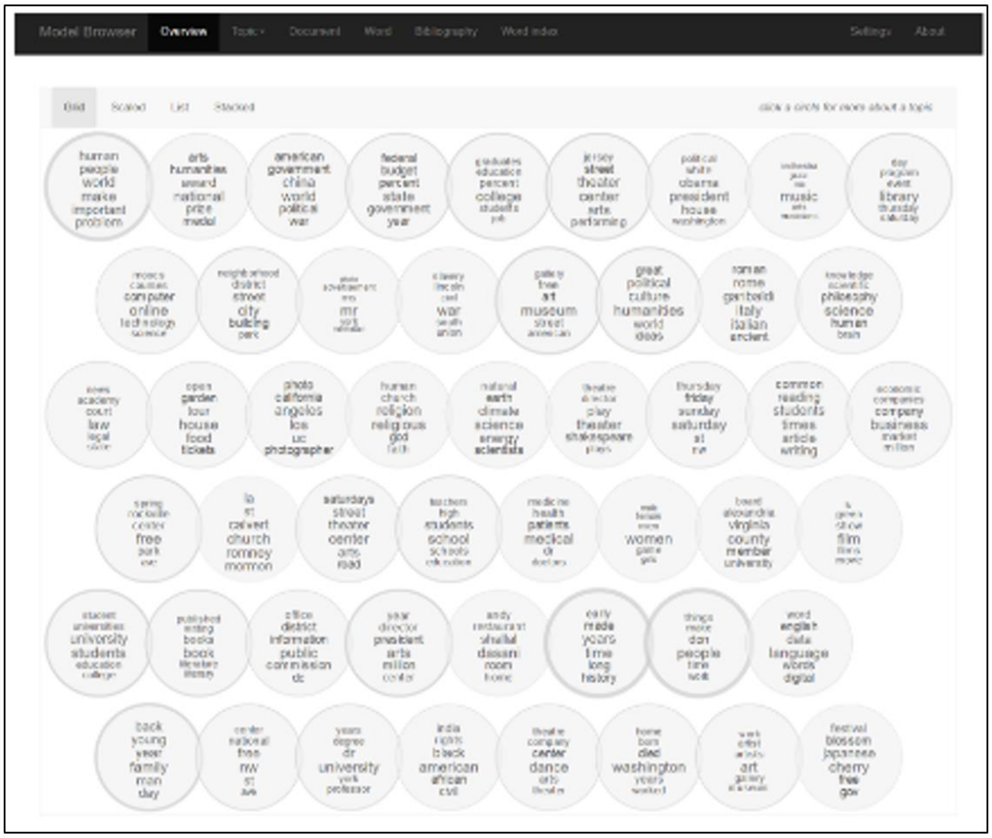

Model Browser
Overview
Topic
Document
Word
Bibliography
Word index
Settings
About
Grid
Scaled
List
Stacked
click a circle for more about a topic
human people world make important problem
arts humanities award national prize medal
american government china world political war
federal budget percent state government year
education percent college students
street theater center arts performing
political obama president house washington
music
day event library
computer online technology science
neighborhood district street city building park
mr
war
art museum street american
great political culture humanities world ideas
roman rome italy italian ancient
knowledge philosophy science human brain
academy court law legal
open garden tour house food tickets
photo california angeles los uc photographer
human church religion religious god faith
natural earth climate science energy scientists
theater director play theater shakespeare plays
thursday friday sunday saturday at
common reading students times article writing
economic companies company business market million
spring center free park
st calvert church romney mormon
saturdays street theater center arts road
teachers high students school schools education
medicine health patients medical dr doctors
women
board alexandria virginia county member university
film
student universities university students education college
published writing books book literature
office district information public commission dc
year director president arts million center
restaurant room home
early made years time long history
things people time
word english data language words digital
back young year family man day
center national free nw st
years degree dr university york professor
india rights black american african civil
theatre company center dance arts
home born died washington years
art gallery
festival blossom japanese cherry free

Figure 2: Dfr-browser interface showing one of several dynamic, interactive views of a topic model generated by WE1S during its pilot project.

[12] See WE1S's comparative study of "Topic Modeling Systems and Interfaces." WE1S adapted Goldstone's dfr-browser, which is open source under the MIT license, with his assistance.

ranked articles associated with a topic, and the changing importance of a topic in the total document set over time.

In addition, WE1S will explore "word embedding" (word2vec) and text-classification analytical methods that have the potential to use the project's collected data in ways that augment topic modeling:

- *Word Embedding*: Whereas topic modeling identifies themes in a corpus of documents based on statistically co-occurring clusters of words, word embedding is a so-called "shallow neural network" approach that mathematically models the semantic relations between a corpus's words themselves, thus making it possible to model not just the co-occurrence of words but the logical relations of analogy and opposition between words. In an often cited example, word embedding thus algorithmically derives such analogies as the following in a corpus of texts: *"king" is to "man" as "queen" is to "woman."* In essence, word embedding holds out the tantalizing promise of showing the relations of conceptual similarity and difference in language that illuminate how a society "thinks" about things through language. It is a computationally tractable method of modeling such modern and contemporary notions of intellectual history as early-20th-century "history of ideas," the mid-20th century *Annales*-school's history of *mentalités*, mid-20th century structuralism, late-20th-century cultural criticism (e.g., Foucault's "epistemes"), and more recent digital humanities "distant reading." (For introductions to word embedding, see Mitra; and Schmidt.)
- *Text Classification*: Machine learning for classification means training a classifier algorithm to predict the probability that a particular text belongs to a pre-determined class. The classifier is trained to recognize the textual features found in a set of documents for which the classes are already known. It is then asked to predict the class for a group of texts that it has not been trained on. Unlike both topic modeling and word embedding, in which collections of words are discovered and conceptual labels applied by researchers after the fact, text classification requires researchers to start with prior assumptions about the classes to which texts belong and then to test the presence of these classes within a given corpus. WE1S plans to use text classification to define the boundaries of “humanities-specific” public discourse by training an algorithm to recognize the textual features of articles that are about the humanities (i.e., articles from its main corpus) as opposed to articles from its random corpus. Stored textual feature data for this algorithm (support vector machines) will be the same kind of non-consumptive, non-reconstructable word lists and word frequency lists as used in topic modeling workflows. (For applications of text classification approaches in literary studies, see Long and So; and Piper.)

f. Technical Methods

WE1S has developed (at a beta stage in the pilot project described in *section II.a below*) a technical environment that implements its research through methods for (1) corpus assembly and preparation; (2) data provenance and workflow management; and (3) the integrated, containerized operation of workflows (including topic modeling and visualization of results). It is also developing (4) a protocol for interpreting topic models that lays out in declared form the iterative steps of interaction between human interpreters and machine-learning results. (This last item is currently still in progress.) While particular features of this technical environment are specifically customized for WE1S, the overall paradigm is generalizable to many other digital humanities projects and can be implemented either "as is" through WE1S's open-source methods and tools or by adapting these.

In detail, the elements of the WE1S technical environment in its beta form are as follows (see *"Glossary for WE1S Technical Environment"* in the *Appendix* for definitions of terms):

1. Corpus Assembly and Preparation System

As described under "Research Methods" above, WE1S collects as "plain text" (the most tractable format for computational analysis) the materials from its journalistic sources via such databases as ProQuest and LexisNexis (or directly through the APIs of source publications). In the case of databases, it does so by first using manual means for searching and downloading (as dictated by licensing conditions), and secondly using automated means for "scraping" (as plain text), cleaning, and other pre-processing of downloaded documents into "bags-of-words" analytical data. The cleaning, pre-processing, and conversion of plain texts into "bags of words" occurs in the project's Virtual Workspace System and in a secure annex of that system (see below), which then also mounts all or various parts of the WE1S dataset for topic modeling (and also de-duplicates material such as data from articles collected twice because they contain both "humanities" and "liberal arts," which WE1S searches on separately).

2. Manifest Framework

Digital humanities researchers working with large data sets or iterative processes have in the past adopted localized, ad hoc means for keeping track of their data, processing steps, and results--an approach that impedes collaborative work, makes repeating or adjusting research processes difficult, and does not support emerging publication standards for transparent data provenance and reproducible research (as in the data sharing policy of the new *Cultural Analytics* journal).

WE1S addresses these issues through a "manifest framework" that documents the components and relations between different parts of a digital humanities research workflow--including data collection, pre-processing (e.g., cleaning), analysis, and presentation (e.g., visualizations). The WE1S manifest framework consists of a generalizable method for annotating data provenance and workflow declared through schema-based documents known as "manifests":

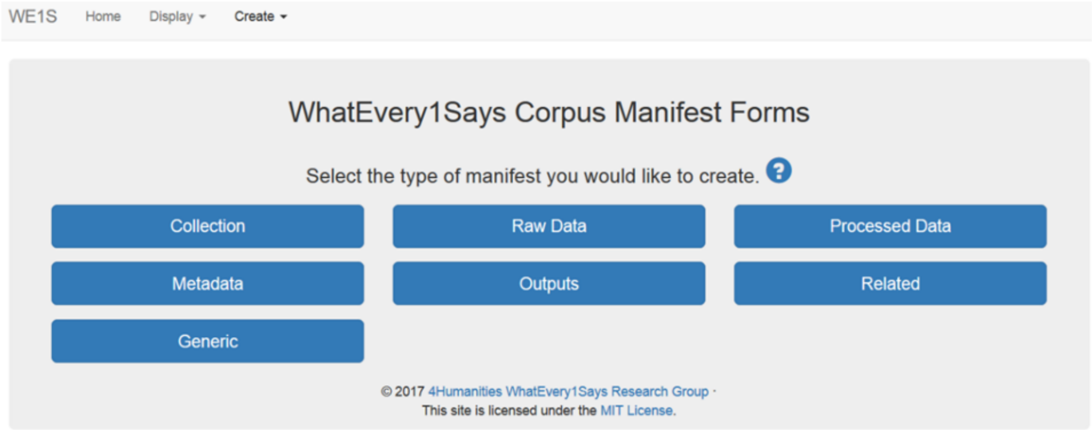
- A manifest is a plain-text file formatted in the JSON serialization format (organizing information into keyword-value pairs) for describing a resource or process. Manifests can be used for a variety of purposes, but their primary intent is to help humans document and keep track of their workflow. The JSON format employed by WE1S is thus generally readable by humans. However, it can also be parsed by computers, thus allowing manifests to function as configuration files for scripts and digital tools in the WE1S Virtual Workspace System. Manifests can be created in a simple text editor. They can also be written in other formats such as XML or YAML, and can be converted to JSON (or vice versa) as appropriate to particular projects.

- Manifests conform to the WE1S manifest "schema" (a definition of the terms and logic needed for tracking WE1S resources and processes that is currently in a version 1.0 state). This schema declares the required and optional properties needed to document different kinds of resources or processes in the project.[13] The inspiration for this approach comes from well-established standards used in the digital humanities--e.g., the Text Encoding Initiative (TEI) and the International Image Interoperability Framework (IIIF). While the WE1S schema is by design simpler than these standards, it can be expanded and customized based on the needs of other types of projects. The schema itself is encoded using the JSON Schema specification for defining the structure of documents containing JSON data, which readily allows for validation and adaptability as well as unified storage with the workflows of the Virtual Workspace System (see below), which are also JSON.
- The WE1S manifest framework shares much in common with other metadata standards and workflow management tools deployed in the sciences and other fields (such as the W3C's PROV Ontology and the Open Science Framework). But for use in the humanities it is designed specifically around a schema suited to the kinds of materials and processes typical of humanities research and also requires relatively little technical overhead. The WE1S manifest schema is also extensible, allowing for project customization. And there are many tools in common programming languages for validating manifests against the manifest schema. Since manifests are text documents, they are easy to adopt by individual scholars working on small projects. For larger collaborative research projects, the WE1S Workflow Management System can scale up the manifest framework (with the WE1S project itself serving as proof of concept.) Content in the manifest framework can also be cross-walked to other tools or metadata standards as needed.

[13] Documentation of the WE1S manifest schema version 1.0 is available at https://github.com/scottkleinman/WE1S/blob/master/WE1S-Schema-1.0.md.

3. Workflow Management System

The Workflow Management System is a Web-based platform for creating and managing manifest documents (see Figure 3). It allows researchers at various levels of technical proficiency to create valid manifests by filling in forms in their browser. Users enter manifest information required by the WE1S schema in Web-based forms. Alternatively, the platform can import manifests to the database from pre-existing manifest documents. The Workflow Management System is particularly important for newcomers to the WE1S project (e.g., new research assistants) who may not be familiar enough with the WE1S schema to create valid manifests from scratch. It also provides the ability to search the project's stored manifests, which will become the basis for part of the public-facing Web site at the end of the project.



WhatEvery1Says Corpus Manifest Forms

Select the type of manifest you would like to create. ?

© 2017 4Humanities WhatEvery1Says Research Group ·
This site is licensed under the MIT License.

Figure 3: Web interface for WE1S manifest system.

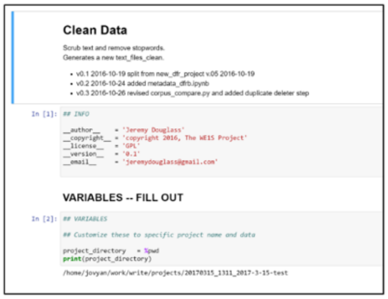
The Workflow Management System is built on a lightweight Python-Flask Web framework that is uncomplicated to deploy. Although manifests do not require a database storage system and can be used independently in flat file format, the Workflow Management System is backed by a MongoDB database, a contemporary "non-relational" (NoSQL) database that stores records in a JSON-like format similar to a manifest file. The system generates Web-based forms automatically from the WE1S manifest schema using the open source Javascript library Alpaca Forms. This means that changes to the schema automatically update the forms and database structure in the Workflow Management System, making the system adaptable as a project evolves.

4. Virtual Workspace System

To address a range of computing demands from a geographically distributed team with varying technical skills and different workstations, WE1S has created a Virtual Workspace System that facilitates open, reproducible digital humanities research through a defined computing platform, a shareable online environment, integrated customizable workflows, and on-demand online presentation of results. The WE1S Virtual Workspace System runs through the Web as well as locally on a laptop; its design and implementation can be used by other digital humanities projects; and it is consonant with the philosophy of such other online or containerized integrated systems as Lexos or DH Box that make advanced digital humanities research environments accessible.

Specifically:

- The WE1S Virtual Workspace System is a virtual environment (runnable online from a server or as a "containerized" virtual computer on a local workstation) that implements a computing platform and a directed series of workflows. Data workflows include those for cleaning and pre-processing texts; converting texts into non-consumptive use "bags of words"; selecting parts of the WE1S dataset to analyze; generating topic models; and outputting results. These workflows are implemented using Jupyter notebooks (previously known as "iPython notebooks"), which both document processing steps and run actual code in step-by-step modules. Such notebooks are a powerful tool for the digital humanities because they guide users at various levels of programming fluency through data procedures--doing so either automatically ("run all") or with decision power at necessary points (e.g., inputting how many topics to ask for in a topic model).



Clean Data

Scrub text and remove stopwords.
Generates a new text_files_clean.

- v0.1 2016-10-19 split from new_dfr_project v.05 2016-10-19
- v0.2 2016-10-24 added metadata_dfrb.ipynb
- v0.3 2016-10-26 revised corpus_compare.py and added duplicate deleter step

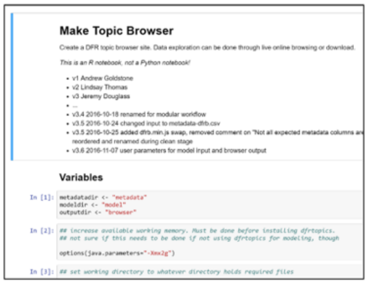
```
In [1]: ## INFO

        __author__    = 'Jeremy Douglass'
        __copyright__ = 'copyright 2016, The WE1S Project'
        __license__   = 'GPL'
        __version__   = '0.1'
        __email__     = 'jeremydouglass@gmail.com'
```

VARIABLES -- FILL OUT

```
In [2]: ## VARIABLES

        ## Customize these to specific project name and data

        project_directory   = %pwd
        print(project_directory)

        /home/jovyan/work/write/projects/20170315_1311_2017-3-15-test
```

Figure 4: Jupyter data notebook for cleaning texts before topic modeling.

- An effective innovation of the WE1S Virtual Workspace System is that its Jupyter notebooks are chained in series. This means that a workflow enacted by one notebook automatically calls the next workflow in a logical sequence. For instance, a humanities user who runs the notebook for cleaning (and other pre-processing) steps on the WE1S corpus is led at the end of the process to a notebook for topic modeling the materials. Another innovation is that workflows in WE1S are all based on a project template system. Each new project begins by generating a new project folder containing a copy of a chained set of default notebooks. The project can then be customized. For data exploration, this setup encourages researchers to create dozens of related project explorations -- each encapsulating their own code, configurations, and metadata – rather than constantly tweaking a single project. A successful run then stands as a record, and may be archived.



Make Topic Browser

Create a DFR topic browser site. Data exploration can be done through live online browsing or download.

This is an R notebook, not a Python notebook!

- v1 Andrew Goldstone
- v2 Lindsay Thomas
- v3 Jeremy Douglass
- ...
- v3.4 2016-10-18 renamed for modular workflow
- v3.5 2016-10-24 changed input to metadata-dfrb.csv
- v3.5 2016-10-25 added dfrb.min.js swap, removed comment on "Not all expected metadata columns ar reordered and renamed during clean stage
- v3.6 2016-11-07 user parameters for model input and browser output

Variables

```
In [1]: metadatadir <- "metadata"
        modeldir <- "model"
        outputdir <- "browser"

In [2]: ## increase available working memory. Must be done before installing dfrtopics.
        ## not sure if this needs to be done if not using dfrtopics for modeling, though

        options(java.parameters="-Xmx2g")

In [3]: ## set working directory to whatever directory holds required files
```

Figure 5: Jupyter data notebook for visualizing a topic model using dfr-browser.

- A third innovation is that the final Jupyter notebook in the WE1S Virtual Workspace System generates an on-the-fly Web site showing a dynamic, interactive view of a topic model in Andrew Goldstone's dfr-browser interface. This Web site can be automatically and iteratively recreated whenever the underlying data workflow is changed (e.g., when a workflow is repeated using different parts of the corpus or different topic-modeling parameters).

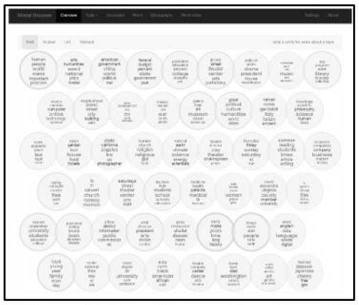


Figure 6: Resulting online view of dfr-browser's dynamic interface for exploring topic models.

5. Interpretation Protocol for Topic Models

Because complex data-analysis sequences can have a "black box" effect, one of the needs of current *in silico* science is not just to document technical workflows for reproducibility but also to make humanly understandable the steps in a workflow. The goal is to facilitate the interpretation of results. For example, a recent paper by Yolanda Gil and Daniel Garijo titled "Towards Automating Data Narratives" provides proof-of-concept for the automatic creation of prose "narratives" of data workflows from the steps recorded in the Wings workflow system. An example of such machine-generated explanation quoted in their paper is as follows:

> The topic modeling method has five steps. The first one, Stop words step, uses an input dataset and a words dataset to produce a filtered result. Next, the Small words step consumes that output to produce another filtered result. The next step is the Format dataset a reformatting step which adapts the result for the Train topics step. Next, the TrainTopics step produces an output topics dataset. Finally the Plot topics step is [*sic*] takes the output topics dataset to create the term-topic matrix visualization.

Digital humanities research, of course, is rooted not just in data science but also long-standing traditions of humanistic hermeneutics, including the critical scrutiny of how humans "read" and "interpret" materials. Digital humanists thus carry the extra burden of needing to make visible the machine-to-human and human-to-human interpretive steps hidden in such scientific narrations of process as the one instanced above--for example, steps involving how researchers read a topic model and how researchers communicate, discuss, and provide evidence for observations about topic models to reach credible conclusions. Yet there are currently no best practices in the digital humanities for explaining data workflow, let alone with attention to the act of human interpretation. In the case of topic modeling in the digital humanities, for example, there are few studies that provide transparent descriptions of the interpretive assumptions, steps, and iterations needed to decide how many topics to seek, what topics are interesting, how the topic model guides the researcher back to specific articles for examination (and vice versa), and how groups of researchers collaborate in using a topic model to generate hypotheses or come to conclusions.

As part of its technical methods, WE1S is developing a topic-model interpretation protocol that declares in understandable form (as part of a manifest workflow) step-by-step interactions between machine learning and researcher interpretation/collaboration (e.g., when in the process researchers convene to interpret a topic model; what outputs, visualizations, and secondary algorithmic products such as Principal Component Analysis or hierarchical clusterings are used to deduce groups of topics; how researchers discuss a topic model; and how topic models and interpretive acts are iterated). The goal is not to assert *the* definitive topic-model interpretation process (because this will be different depending on the nature of a project, its materials, and its personnel), but to declare *a* topic model interpretation process that can then serve as a model and be adapted, improved, and varied by the larger DH community. It may be that over time one or several kinds of digital-humanities data interpretation protocols will evolve as shared conventions.

While still in progress, the WE1S topic-model interpretation protocol is currently drafted to specify a sequence of interpretive steps as follows (showing only high-level steps). These steps will eventually be documented in JSON-formatted manifests according to the WE1S manifest framework. This will allow for provenance tracking of interpretation workflows, the sharing of such workflows with other projects, and the automation of steps that facilitate interpretation--e.g., allowing the WE1S Virtual Workspace System to generate visualization aids at appropriate moments as the interpretation process unfolds).

- *Interpretation Stage 1*
    - Automatic generation of topic models at various scales of granularity and with different parameters.
    - Automatic and manual creation of materials to facilitate interpretation of topic models (e.g., visualizations of topics, classifications and cluster analyses of topics)
    - Initial researcher assessment of topic models by a team working according to a checklist of steps defined for the WE1S project but adaptable with variation to other projects. (For example, given the nature of the newspaper articles that are its primary material, WE1S will work out an optimal way to conduct initial assessment of the usability of a topic model by inspecting topics flagged for attention as not immediately recognizable or comprehensible, e.g., by reading samples of the top articles contributing to those topics.)
- *Interpretation Stage 2*
    - Reiterated automatic generation of topic models for optimization based on initial researcher assessment of preliminary results.
    - Detailed researcher analysis of topic models (by a team working according to a checklist of steps that reiterate those mentioned above but also add such extra steps as reading more sample articles, comparing articles that contain a high proportion of the same topics, or using statistical clustering aids such as the hierarchical dendrogram visualizations produced in Lexos to group topics).
    - Study of major topics and clusters of topics.
    - Write-up of analyses and observations in a standard format.
- *Interpretation Stage 3*
    - Researcher comparative analysis (machine-assisted) of correlations/differences between various parts of the corpus as they appear in the topic model (e.g., proportional weight of specific topics in one set of newspapers versus another).

    - Researcher comparative analysis (machine-assisted) of temporal trends in topics (e.g., comparing the 1980s and the 2010s).
- *Interpretation Stage 4*
    - Detailed analytical and interpretive reports written in a standard format.
    - Creation of “data sheets” (samples of evidence, links to various visualizations of the topic model, and notes on observations) to support reports.

II. Plans for Expansion from Pilot Project

a. Original Pilot Project

WE1S has operated since 2013 as a pilot project supported by small UCSB faculty research grants (a total of $19,800 awarded to PI Liu so far). In this phase, the project focused on developing research questions and goals, research methods, and technical methods *(as described in sections I.b-f)*. It also created partial, experimental topic models (such as seen in Figure 1) based on data for the last 20 to 30 years from a small set of English-language newspapers and other journalistic sources.[14] The pilot project also gathered a small "random" sample of articles from its sources *(described in section I.e)*. To date, there is no public-facing project Web site except for a short description at http://4humanities.org/category/whatevery1says/. The developers' Web site for the pilot project, which hosts extensive planning documents, progress reports, and meeting notes, is at http://4humwhatevery1says.pbworks.com.

Since the main purpose of the pilot project has been to develop goals, methods, and technical implementation based on relatively small samples of materials (and since the WE1S faculty have been constrained in their ability to work steadily on the project due to limited funding, research assistance, and time on top of their regular teaching, research, and administrative duties), there are not yet conclusive results. However, analyses of experimental topic models produced along the way--especially during three iterative, team-wide seminars on models conducted in November 2015, January 2016, and November 2016--have shown the promise of the topic modeling approach. It was during the course of these analyses, for instance, that WE1S discovered that there is a foreground / background issue to be studied in understanding the impact of the humanities on public life. At times, the humanities surface as focal areas of concern (as in discourse on the "humanities crisis"). At other times--and, indeed, in most times--the humanities are discussed as part of the backdrop, baseline, or immersive medium of personal, social, political, and other collective life (as in the large number of wedding announcements, obituaries, and event listings that mention the humanities). It was in the course of wrestling with such miscellaneous materials that WE1S came to the realization that they are not a "bug" but a "feature" of the research problem--one now formulated as the first of the project's research hypotheses mentioned in *section I.b* above--*That newspaper articles and other documents containing the literal phrases "humanities," "liberal arts," and "the arts" are*

[14] Sources included the following newspapers: *The Guardian* (London), *The Globe and Mail* (Toronto), *The Los Angeles Times*, *The New York Times*, *The Wall Street Journal*, *The Washington Post*, and USA Today. They also included one magazine, *The New Yorker*, and transcripts from PBS broadcasts.

likely places to look for focused discussion of the humanities (e.g., articles on the "humanities crisis") and socially broad discussion of the humanities (e.g., articles on the humanities as part of personal life and general culture). Because WE1S makes such findings quantitatively tractable--inviting inquiry, for example, into which foreground themes and background contexts are prevalent, and at what level of intensity--new avenues are opened for understanding, and advocating, the humanities in their complex entanglement with modern society.

b. Mellon-Funded Next Stage of Project

If Mellon funding is granted, WE1S will during a three-year timeline (beginning October 1, 2017) greatly extend the scope and diversity of its sampled materials from public discourse; improve its research methods and technical implementation to enable more rapid and more flexible exploration of these materials; and produce analyses and other outcomes for its intended public, humanities, and digital humanities audiences. The primary aim of WE1S's next stage is to jump the project to a higher level of scholarly significance (in which research conclusions will be based on more representative materials and open methods and tools) and also greater potential public impact (e.g., analyses of public discussion of the humanities and recommendations for humanities advocacy) *(see section II.f, Expected Outcomes and Benefits below)*.

A timeline of three years is proposed so that the first year can be devoted to scaling up to a more representative corpus of articles and other documents and to improving methods for collecting and analyzing it. In addition, such an initial "scaling up" year will allow WE1S to convene an advisory group of digital humanities experts (and also some domain experts in fields such as race and ethnic studies relevant to the expanded scope of the project) to offer early feedback as the project jumps to a larger scale with more ambitious methodological and analytical aims.

Specifically, WE1S's plans for the three years of a Mellon grant are as follows:

1. Corpus Expansion

WE1S will expand the range and representativeness of its primary corpus of contemporary journalistic publications (defined as newspapers, magazines, and radio/TV transcripts of news or talk shows available in English across multiple nations). Through institutional subscriptions to commercial databases--e.g., LexisNexis Academic, ProQuest (including News and Newspapers, ProQuest Historical Newspapers, ProQuest's ethnic, race, and gender news databases)--WE1S's researchers have access to over 2,500 English-language newspapers from which full-text digital articles of the past few decades can be collected (through a combination of manual and automated means conforming to source licensing terms) and converted for text analysis operations into non-consumptive-use datasets.

WE1S plans to devote research at the beginning of its timeline to determine which specific sources to target in these areas that will be most representative and useful for the project's goals. While the criteria for representativeness and usefulness will evolve iteratively as the project team begins its research on potential sources *(see under Project Year 1 in section II.e, Activities and Timeline)*, WE1S has initially identified two key areas for corpus expansion:

- The first is the geographical and national scope of its corpus of materials: WE1S will investigate expanding the range of its sources by including materials from Anglophone newspapers located outside North America. Such newspapers include *The Times*, *The Sunday Times*, and *The Independent* in the United Kingdom; *The Australian* and *The Daily Telegraph* in Australia; *The New Zealand Herald* in New Zealand; and *The Times of India* in India. Initial criteria for inclusion include a publication's value for representing a part of the world not previously included; its national or regional circulation; and the technical feasibility of collecting and processing its articles. WE1S will also draw on current research on media impact to help it develop a strategic rationale for selection of materials (e.g., the approaches to defining and measuring the impact of journalistic media surveyed by Schiffrin and Zuckerman).
- The second area for corpus expansion concerns what may be called the social scope of WE1S's materials. An especially high-priority goal is to include sources that can allow WE1S to ask research questions about how the humanities are viewed by, or in relation to, different social groups (racial, ethnic, gender, immigrant, and age). This is a diversity aim that is organic to WE1S's core research. Because both historical and contemporary anecdotal evidence suggests that particular groups channel themselves (or are channeled) into career choices that make the humanities a lesser priority during first-to-college or first-generation-immigrant stages in their social trajectory, WE1S hypothesizes that researching "what everyone says about the humanities" in particular groups can add meaningfully to society's more common talking points about numbers of humanities majors, career goals, or the relation of the humanities to the sciences or business. To facilitate such research, WE1S will include in its primary corpus journalistic materials provided by databases such as Ethnic Newswatch, Proquest Black Newspapers, and Proquest U.S. Hispanic Newsstand. These are the sources for this purpose that WE1S has so far identified from canvassing the databases available to its researchers through institutional subscription and also from initial consultation with scholars and university administrators working in race and ethnic studies. WE1S will seek further resources. If feasible, WE1S will also attempt a small-scale experiment in topic-modeling a limited sample of articles from Spanish-language newspapers, though existing topic modeling and other text analysis methods are not capable of integrating multilingual materials in the same model. Criteria for inclusion of materials in WE1S's research corpus will be a source's value for representing part of the "social scope" of the humanities not previously included, the publication’s circulation and intended audience, and the technical feasibility of collecting its articles.

In addition to expanding its primary corpus of materials as outlined above, WE1S plans to extend the range of research questions it can pose by collecting smaller "sub-corpora" of other kinds of sources that can be folded into, or separated from, its main corpus as needed for computational analysis. Particular sub-corpora will be chosen after detailed research at the beginning of the project timeline. Steps in such research will involve consulting scholars and university administrators as well as WE1S's advisory board; reading and discussion of sample materials from potential sub-corpora; assessment of technical feasibility (i.e., can a source be used in a way that fits practically into the project's technical workflow); and assessment of strategic value (e.g., does a sub-corpus add meaningfully to the representativeness of the project's materials or provide needed perspective on questions that emerge in analysis of previously gathered materials). Sub-corpora are likely to include some of the following:

- Historical newspaper coverage of the humanities from earlier in the 20th century (gathered through ProQuest Historical Newspapers; the Library of Congress's Chronicling America resource; and, in some cases, through the archives and API's of individual newspapers);
- Government and political documents (gathered through resources such as Congress.gov, Whitehouse.gov, U. S. Government Publishing Office, and the archives of individual states, with data gathering assisted by API's from the Sunlight Foundation)[15];
- Reports and publications by scholarly and professional associations as well as grant agencies and foundations[16];
- Public documents of higher-education institutions that mention the humanities (e.g., so called university "viewbooks"; mission statements of humanities centers; and speeches by campus presidents and deans);
- Scholarly research articles discussing the humanities (collected from JSTOR). A particularly rich avenue of research will be to use the recently introduced JSTOR Labs Text Analyzer service to discover research articles relevant to sample materials from the WE1S corpus. (Because the Text Analyzer builds on JSTOR Lab's own usage of topic modeling, there may also be ways that WE1S can use Text Analyzer to corroborate or extend WE1S analyses of topic models.)

To guard against "mission creep" (e.g., continuously adding sub-corpora to pursue multiplying research questions, or attempting to gather sub-corpora that require the development of new methods of manual and automated harvesting), WE1S will conduct a triage assessment at the end of its second development year *(see section II.e, Activities and Timeline)* to ensure that it focuses on high-value sub-corpora whose collection and analysis can be finished within the project's timeline.

[15] For WE1S's preliminary scoping study of U.S. Congress, White House, and selected state documents related to the humanities, see 4Humanities.org, "What U.S. Politicians Say About the Humanities--A Data Set and Analysis."

[16] An example is the 2013 report titled *The Heart of the Matter* from the American Academy of Arts & Sciences' Commission on the Humanities and Social Sciences. For WE1S's topic-model study of this document, see 4Humanites.org, "The Heart of the Matter Topic-Modeled (A Preliminary Experiment)."

2. Improvement of Research and Technical Methods

Currently, WE1S's research methods and their technological implementation *(see sections I.e and I.f above)* are first-generation. The research apparatus WE1S created for its pilot project works adequately to allow a distributed group of researchers and assistants with various levels of technical expertise to collaborate in staging, managing, and tracking the movement of textual materials through analytical and modeling processes into a dynamic, visualized interface for interpretive exploration. The limitation of WE1S's current system, however, is that it is slow and labor-intensive at the initial step of ingesting materials (scraping plain text from a variety of sources with different methods depending on source); constrained to a single analytical method in its middle steps (topic modeling); and constrained to one kind of exploratory interface in its end steps (dfr-browser). Additionally, WE1S has encountered problems in generating high-dimension topic models (over 300 topics) due to hardware and software constraints. To support the aim of flexibly and rapidly asking research questions about a larger, more diverse corpus of materials, WE1S plans to evolve its technological research environment.

For this purpose, the programming expertise of two of its co-PIs, Jeremy Douglass and Scott Kleinman, will be supplemented by that of research assistants from talent pools the project has drawn on in the past, such as students in UCSB's Computer Science department and Media Arts & Technology program and their equivalents at CSUN. WE1S also anticipates needing relatively minor hardware and software upgrades for its development phase. (Development will usually be implemented in-house rather than through third-party cloud or other platforms in order to maintain an agile development process and avoid time-consuming migrations or reconfigurations of the WE1S research environment. However, limited cloud-based “elastic compute” nodes may be used as an alternative to in-house hardware for the purpose of scaling up the operation of algorithms that do not require a full research environment.)

Specifically, WE1S will take up the technical tasks outlined below:

- *Improve methods for quick, iterable corpus assembly* (scraping, cleaning, and other pre-processing of plain text from articles; followed by conversion of text into non-consumptive use data that can be stored). The goal is to allow WE1S researchers rapidly and flexibly to add more materials to the corpus as dictated by evolving research questions. This task involves a variety of scripting and data-preprocessing automation work.

- *Extend search and analytical methods*. Develop additional methods of searching for materials related to the humanities (e.g., corroborating or extending the simpler method of keyword searching through more complex pattern matching), and then study the materials using more than a single method of text analysis. WE1S anticipates exploring "word embedding" (word vector) approaches that could complement topic modeling in the interpretation of complex discursive fields.

- *Improve interfaces for interpretive exploration of results*. Building on dfr-browser (and perhaps also integrating some of the workflow of the Lexos text-analysis system, which WE1S co-PI Scott Kleinman helped develop), WE1S plans to extend the dynamic interface for studying topic models it presently generates through its Virtual Workspace System. For example, WE1S wants to explore how clustering-visualizations and other methods of apprehending relations among topics in a topic model might be used to complement dfr-browser. Improving the interface for interpretive exploration is an important stepping stone toward the eventual public-facing Web site that will represent the project.

- *Improve WE1S's Workflow Management System in conjunction with the project's Virtual Workspace System.* Currently, WE1S has created a prototype for form-based entry into a MongoDB database of provenance and workflow information ("manifests"), and for querying from the database. It has also started integrating its manifest schema with its virtual-machine workspace for workflows. Going forward, WE1S will improve the integration between its Workflow Management System and Virtual Workspace System. This integration will enable users performing a Jupyter notebook procedure in the workspace to call a manifest of information about resources, scripts, and steps to configure that procedure. Reciprocally, it will enable procedures to populate manifests automatically with provenance information serving as the basis for shareable and reproducible workflows. Improving the JSON-based movement of metadata between WE1S's Workflow Management System and Virtual Workspace System is an important stepping stone toward future project dissemination, since it will create the basis for input and output pipelines (similar to APIs) connecting to other researchers' projects.
- *Develop a repository strategy for disseminating and sustaining the project's data and technology*. WE1S's goal is to develop an integrated method for disseminating evolving iterations of its data and technologies, and depositing its materials in a sustainable repository accommodating not just data but the containerized Docker system holding a working copy of the project's virtual-machine workspace. *(For details, see section II.h, Sustainability.)*
- *Develop a public-facing front end for displaying and exploring the results of the project*. The goal is to create a Web site that will show WE1S's results in a dynamic, queryable interface accompanied by explanations and other guidance from the project team. Such a site will encourage the public and scholars to explore public discussion of the humanities in the media, and to read the underlying materials (through links to their proprietary original locations). The interface for the site will draw on output from the project's internal technical systems (its Workflow Management System and Virtual Workspace System). The dynamic presentation of data and topic models will be complemented on the Web site by other kinds of materials, including analyses, recommendations, resources for humanities advocacy, and scholarly deliverables. Also, WE1S will explore the feasibility of using the open-standards based Hypothes.is Web annotation system to add a publicly viewable "layer" of highlights, comments, and

links over selected newspaper articles and other online documents in their original locations, thus creating effective case studies of the project's findings.[17]

c. Diversity and Inclusion

i. In light of its theme, WE1S primarily understands diversity and inclusion to refer to facilitating the ability of underrepresented racial/ethnic groups--and also first-generation-to-college, immigrants, and others--to embrace the humanities in common with others so as to contribute to the full life of individuals, groups, educational and cultural institutions, and, ultimately, society. While there are many exceptions, members of such groups are often seen--as much by themselves as others--to focus disproportionately on a narrow range of disciplinary and career goals during the early trajectory of their group's entry into the social commonwealth. This was the precedent set in the mid-twentieth century when Ivy League universities such as Yale throttled the admission of Jewish students--in part, or ostensibly, because they were perceived to be narrowly focused on scientific and technical fields as opposed to the well-rounded "liberal arts" of the "Yale Man" (see Kabaservice). In different ways, versions of such perceptions about these groups of people have continued, with particular groups identifying or being identified as devoted primarily to STEM, social-science, business, and other non-humanities disciplinary and career goals. Students in these groups find life-fulfilling ways of being and working as "human" that somehow shunt the humanities to the side, even if the humanities is one of their core passions or, in fact, are part of their deep cultural heritage. Alternatively, one of the best entries into the humanities for such students is through the door that is designated by themselves and others as specifically "their" humanities--for example, their particular literature or their history (areas of study provided for by today's robust, innovative majors in such fields).

A core mission of WE1S's plan for expanded research, therefore, is to acquire materials that facilitate understanding the complex relationship of underrepresented and other groups to the humanities. As detailed in *section II.b.1 (Corpus Expansion)*, WE1S has identified resources (databases such as Ethnic Newswatch, Proquest Black Newspapers, and Proquest U.S. Hispanic Newsstand) from which materials can be gathered and folded into its topic modeling analysis that might serve this purpose. In addition, WE1S has begun consulting with scholars and university administrators specializing in the cultures of underrepresented groups to discover more such discursive materials. The aim is to position WE1S to ask such questions as: how do mainstream media position students and others from particular groups relative to the humanities? How do media articles addressed specifically to such groups compare with mainstream media? In what ways does public opinion about the very ideal of "diversity and inclusion" correlate with public opinion about the humanities?[18]

[17] WE1S has consulted with Dan Whaley and Heather Staines, CEO of Hypothes.is about this possibility. In addition, 4Humanities.org initiative is exploring participation in Hypothes.is' new "public groups" annotation communities. As one of the system's public groups whose mission is oriented toward the social good, 4Humanities may be able to help facilitate the creation of a larger community of humanists willing to act as informed, reflective commentators on public discourse.

[18] Because WE1S concentrates on public discourse about the humanities, it does not attempt to conduct ethnographic or sociological research that directly interviews or questions members of particular racial, ethnic, or immigrant groups. Such approaches are important, but they fall outside the project's scope and expertise.

An important context in this regard is that two of WE1S's participating universities--UCSB and CSUN--have been designated Hispanic Serving Institutions (HSI) by the Hispanic Association of Colleges and Universities. UCSB earned this status in 2015 (as well as grants from the U.S. Department of Education's Developing Hispanic Serving Institutions Program). It is the only member of the Association of American Universities (AAU) and just one of a very few R1 universities with this distinction. During the past two decades, UCSB's Chicana/o and Latina/o student population rose from 11% to 26%. Meanwhile, its number of African American, and American Indian/Native American students doubled over the same period as a percentage of undergraduate enrollment. The *New York Times* College Access Index also ranked UCSB No. 3 in 2015 and No. 2 in 2017 among U.S. top higher-education institutions for doing the most for low-income students, based on the proportion of students receiving Pell grants, graduation rates of such students, and tuition levels for low- and middle-income students (see Yang; and *New York Times*, "Top Colleges Doing the Most for the American Dream"). CSUN, one of the 23 campuses in the California State University system, was designated a HSI even earlier in 1997. It is the only four-year institution of higher education serving the San Fernando Valley in Los Angeles, home to nearly 40% of the City of Los Angeles population. Over the past two decades, this locale has seen a shift from a predominantly white majority to a non-white majority. Fifty-one percent of CSUN’s 40,000 students in fall 2016 were from historically underrepresented racial and ethnic backgrounds. Over 46% of CSUN students identify as Latina/o; and another 11% identify as Asian American. (CSUN is also certified as an Asian American and Native American Pacific Islander-Serving Institution.) Additionally, 70% of CSUN's students receive financial assistance as members of low- and middle-income groups; and over one third are the first in their family to attend college (see California State University, Northridge, "Diversity Initiatives"). The third university participating in WE1S, the University of Miami, is a private institution, but has a very strong diversity record as well (see University of Miami, Fact Book 2016-17 (Fall 2016 Fact Book). In Fall 2016, 28% of University of Miami undergraduates identified as Hispanic/Latino, 9% as Black, and 12% as Asian/Pacific Islander. (During the five previous years, University of Miami saw a 3% rise in the proportion of Hispanic/Latino undergraduates, a 19% rise in Black undergraduates, a 4% rise in Asian/Pacific Islander undergraduates, and a 39% rise in students identifying as two or more races.) The University of Miami figures for graduate students are approximately equivalent (29% Hispanic/Latino, 9% Black, 16% Asian/Pacific Islander).

WE1S sees the HSI designation of two of its project institutions and the strong diversity base of all three of its partner institutions as a compelling context for its research interest in the participation of underrepresented groups in the humanities.

ii. Secondarily, WE1S understands diversity and inclusion to refer to facilitating the participation of underrepresented students (and also students such as Asian-Americans underrepresented in the humanities, as well as women underrepresented in technology) in its own project. Because WE1S engages in interdisciplinary humanities/technology research with a strong focus on social issues, an appropriate diversity aim is to include underrepresented students from many fields (humanities, STEM, and social science) in its research as part of their educational training. At the graduate-student level, WE1S has already benefited from diverse participation by research assistants drawn from UCSB's long-established digital humanities emphases in the English

Department, Film & Media Studies Department, Comparative Literature Program, and Media Arts & Technology Program. Core research assistants for WE1S have thus included women and people of color. And at the undergraduate level, one of the defining challenges and successes for WE1S in regard to diversity has been the "Making the Humanities Public" project it is running at UCSB in academic year 2016-2017 (in progress at the time of the writing of this grant proposal). Directed by the WE1S Principal Investigator (an Asian-American), and co-directed by a WE1S graduate-student research assistant (an African-American), "Making the Humanities Public" is an undergraduate collaborative research group funded by a UCSB English Department alumni endowment (The John and Jody Arnhold Endowment). The group consists of ten honors-level undergraduate English majors who are highly diverse (including students of Chicana/o and Asian-American ancestry and a majority of women). This research group spent their Winter quarter (January-March 2017) studying a portion of the WE1S pilot project's newspaper corpus with the aid of a WE1S topic model. They produced a white paper analyzing the way the humanities appear in public discourse and wrote up recommendations on representing the humanities to the public. In spring quarter 2017, they are creating digital humanities projects implementing their recommendations--e.g., making infographics on the humanities based on their surveys of fellow students; posters and graphics showing that "The Humanities Are Closer Than You Think" (the title of a project showing the humanities background of famous public figures); and storymaps on the intersection of humanities and STEM fields. They are also designing a social media and publicity campaign to disseminate their work.

This student "Making the Humanities Public" research group is an indicator of how WE1S plans to include diverse students in its expanded work plan. WE1S's Summer Research Camps will be a vehicle for this purpose *(see description in section II.e, Activities and Timeline)*. At UCSB, these will be advertised to graduate-student research assistants from diverse student bases, including not just programs with strengths in the digital humanities but also in areas such as ethnic and gender studies bearing on the diversity aspect of WE1S's research aims. At CSUN, a particularly important WE1S initiative is to create Summer Research Camps that parallel those held at UCSB but also work to increase campus expertise in the digital humanities, including among its diverse student population. To this end, the CSUN camps will also directly involve undergraduates. In this regard, WE1S hopes to create synergy with CSUN's Mellon-funded HSI Pathways to the Professoriate program, which prepares undergraduate students in the humanities and related fields for careers in the professoriate and works to increase the number of Latina/o professors in the humanities at U.S. colleges and universities. WE1S has begun consulting campus leaders of the HSI Pathways to the Professoriate program, and will recruit student participants from within the program to its summer research activities.

d. Staffing

WE1S plans to conduct its research with the following staff:

- *Project Principals*. The principals will work throughout the academic years and during the summer research camps on project design, management, research, output, and dissemination, as well as on supervision of research assistants. While each principal will be engaged in all aspects and phases of research (in various research, supervisory, or consultative ways), each PI will also have individual leadership and managerial roles for specific areas of the project and for the tasks and task groups associated with those. The nature of these roles and tasks are defined in detail in *section II.e, Activities and Timeline*. In general, however, the areas of responsibility of the principals are as follows:

PI Alan Liu will supervise the project as a whole--taking the lead in setting project directions, recruiting postdocs and RAs at UCSB, coordinating work among the project team (principals, postdocs, and research assistants), organizing and running all-project-participant meetings, managing the work timeline, overseeing and approving budget expenditures, and resolving task conflicts or unexpected challenges. He will also lead or co-lead some specific tasks *(as described in section II.e)*. With co-PIs Douglass and Thomas, he will supervise the summer research camps for RAs at UCSB.

Co-PI Jeremy Douglass will take the lead in developing WE1S's Virtual Workspace System. This involves working with RAs to: improve the data collection, analysis, and processing functions of the system; deploy secure methods for the proper handling of project data; develop and test software in the system; and integrate the system with the WE1S Workflow Management System. Douglass will also lead RAs in experimenting with additional searching methods and text-analysis methods. With co-PI Kleinman, he will lead the development of the WE1S public Web site. Also with co-PI Kleinman, and in consultation with Thomas Padilla, he will plan and implement the WE1S repository system for sustainability. And with PI Liu and co-PI Thomas, he will supervise the summer research camps for RAs at UCSB.

Co-PI Scott Kleinman will take the lead in developing the WE1S manifest framework and Workflow Management System, including software development and testing as well as integration with the Virtual Workspace System. With co-PI Douglass, he will lead the development of the WE1S public Web site. With co-PI Thomas, he will customize the dfr-browser interface for exploring topic models so that it can use metadata provided by the WE1S manifest framework. With co-PI Douglass, and in consultation with Thomas Padilla, he will plan and implement the WE1S repository system for sustainability. He will also hire RAs for the summer research camps at CSUN and, with Professor Mauro Carassai, supervise them.

Co-PI Lindsay Thomas will take the lead in developing the WE1S "random" comparison corpus; and will be co-lead with PI Liu on the task of corpus expansion and collection. She will also hire/supervise research assistants at U. Miami. With her RAs, she will explore the use of such analytical methods as classification algorithms to complement WE1S's topic modeling work. With co-PI Kleinman, she will customize the dfr-browser interface for exploring topic models so that it can use metadata provided by the WE1S manifest framework. Together with PI Liu and co-PI Douglass, she will supervise the summer research camps for RAs at UCSB. (She will be in residence in Santa Barbara for the research camps.)

Short bios for the principals are below, and the *Appendix* provides a full curriculum vitae for each.

- *Principal* Investigator *(Alan Liu, UCSB)*--
  Alan Liu is Professor in the English Department at the University of California, Santa

Barbara. He has published books titled *Wordsworth: The Sense of History* (1989); *The Laws of Cool: Knowledge Work and the Culture of Information* (2004); and *Local Transcendence: Essays on Postmodern Historicism and the Database* (2008). His new book, *Friending the Past: The Sense of History in the Digital Age*, is forthcoming from University of Chicago Press. Recent essays include "Hacking the Voice of the Shuttle: The Growth and Death of a Boundary Object" (2016), "Is Digital Humanities a Field?—An Answer from the Point of View of Language" (2016), "N + 1: A Plea for Cross-Domain Data in the Digital Humanities" (2016), "The Big Bang of Online Reading" (2014), "The Meaning of the Digital Humanities" (2013), and "Where is Cultural Criticism in the Digital Humanities?" (2012). Liu started the Voice of the Shuttle Web site for humanities research in 1994. Funded projects he has led as PI include the University of California Transliteracies Project on online reading and the RoSE (Research-oriented Social Environment) software project. Liu is founder and co-leader of the 4Humanities.org advocacy initiative.

- *Co-PI (Jeremy Douglass, UCSB)--*
  Jeremy Douglass is an Assistant Professor of English at University of California, Santa Barbara. He has served for five years as the faculty director of Transcriptions, a center for research in literature, culture, media, and the digital humanities. He also serves as faculty director of the UCSB Digital Arts and Humanities Commons. He is co-author, with Jessica Pressman and Mark C. Marino, of the book *Reading Project: A Collaborative Analysis of William Poundstone's Project for Tachistoscope {Bottomless Pit}* (Iowa UP, 2015), and co-author, with Montfort et al., of the book *10 PRINT CHR$(205.5+RND(1)); : GOTO 10* (The MIT Press, 2012). He recently published the article “Numeracy and electronic poetry” (2015). Douglass currently conducts research on interactive narrative, electronic poetry, and games, with a particular focus on applying the methods of software studies, critical code studies, and information visualization to the analysis of digital texts. His work has been supported by agencies including the NEH Office of Digital Humanities, MacArthur Foundation, Mellon Foundation, ACLS, Calit2, HASTAC, and NERSC.

- *Co-PI (Scott Kleinman, CSUN)--*
  Scott Kleinman is Professor of English and Director of the Center for the Digital Humanities at California State University, Northridge. His 1997 dissertation from Cambridge University focused on Old English phonology and used corpus-based text analysis to examine metrical patterns in early English. His publications on Laȝamon's *Brut* and *Havelok the Dane* focus on regional and legal cultures in medieval England. Kleinman is Co-Director of the NEH-funded Archive of Early Middle English project, which is producing a hybrid digital edition/archive of the surviving manuscripts containing Middle English between 1066 and 1350. He also co-directs the NEH-funded Lexomics project, which produces the Lexos text analysis software. He is both a designer/developer for Lexos and an active investigator of the use of text analysis techniques, particularly for the study of ancient languages. Most recently, he co-authored “Modeling the contested relationship between Analects, Mencius, and Xunzi: Preliminary evidence from a Machine-Learning Approach,” a topic modeling-

based study of Classical Chinese literature forthcoming in the *Journal of Asian Studies*. Kleinman was previously a member of the One Week | One Tool team that produced Serendip-o-matic, a search engine for large open-access image databases. In June 2017, Kleinman was a Fulbright Specialist in Nepal, helping institutions there adopt Digital Humanities in their teaching and research. He has experience working with a variety of Web technologies, including the Python-Flask-MongoDB stack used in the WhatEvery1Says research technical environment.

  - *Co-PI (Lindsay Thomas, U. Miami)--*
    Lindsay Thomas is Assistant Professor of English at the University of Miami. Her research focuses on contemporary US literature, cultural and media studies, and the digital humanities. Her current book project, *Training for Catastrophe: National Security and the Management of the Future*, investigates the use of fiction as a mode of knowledge production within contemporary U.S. national security discourse, arguing that this dependence on fiction trains us to accept catastrophe as part of everyday life. Her work has appeared or is forthcoming in *Surveillance & Society*, *American Literature*, and the edited collections *American Literature in Transition: 2000-2010* and T*he Routledge Companion to Risk and Media*. She has worked as a member of the RoSE (Research-oriented Social Environment) software project and for the 4humanities.org advocacy initiative.

- *Sustainability and Usability Advisor (Thomas Padilla, UCSB)--*
  Appointed in 2016 on the staff of the UCSB Library as the campus's (and one of the world's) first designated Humanities Data Curators, Thomas Padilla will be committing 10% of his effort (approved by the Library) as an embedded member of the WE1S project. Padilla's former positions include Digital Scholarship Librarian at Michigan State Universities, and Scholarly Commons Assistant and Digital Preservation Research Assistant at University of Illinois, Urbana Champaign. His publications include articles on digital preservation and humanities data; and he is the PI of the "Always Already Computational: Collections as Data" initiative, funded by a grant from the Institute of Museum and Library Services (IMLS). Padilla will use his combined expertise in the digital humanities generally and in library-based curation, repository, and "collections as data" practices to help direct WE1S's data curation strategy and implementation, with additional focus on the usability of WE1S outputs. With the aid of Padilla's planning, WE1S's sustainability strategy (for preserving access to project materials and data as described in *section II.h* below) will be integral to its methodological goal of open, shareable, and reproducible digital humanities research. (For c.v. see *Appendix*.)

- *Two postdoctoral faculty fellows located at UCSB in each of the second and third years of the project timeline (to be recruited through a national/international advertised search).*
  The postdoctoral faculty fellows will each receive compensation in project years 2 and 3, respectively, of $39,235 and $40,020 (plus benefits) sourced from the grant. Additionally, their compensation will be topped off by a further salary of approximately $20,000 (plus benefits) a year from UCSB for teaching up to three courses annually as lecturers, thus bringing total annual compensation up to about $60,000. Two-thirds of the postdoctoral faculty fellows' effort will be devoted to WE1S; and one-third to teaching duties. The

postdoctoral faculty fellows will collaborate with the WE1S PIs in project management and execution, and they will also bring additional skills and research interests that extend the project. In particular, WE1S will seek postdocs whose own research complements, and can benefit from, WE1S's goals and methods (so that their effort on WE1S also furthers their own research). For example: a postdoc might be a digital humanities scholar working in such areas as natural language processing, text analysis, visualization, and social network analysis. Alternatively, a postdoc's research might intersect with WE1S's main intellectual and diversity aims (in such areas as Chicana/o studies, African-American studies, Asian-American studies, or the emerging field of "history of the humanities"). The postdoctoral faculty fellowships are delayed in the WE1S work plan until the second and third years of the project to allow time for recruiting fellows during the first year.

(See *Appendix* for a draft job description for the postdoctoral faculty fellow positions.)

- *Graduate student fellows/researchers ("lead research assistants" on a stipend serving in a project manager role; one each at UCSB and U. Miami).* One graduate student fellow/researcher annually at each of the campuses of UCSB and U. Miami will serve as lead research assistants acting in the role of project managers for their campuses. They will help organize (and also participate in) the project's research activities, coordinate the work of other research assistants, and coordinate work across the project's partner campuses. A project manager is needed at both UCSB and U. Miami because these are the two campuses where research assistants will be working on corpus collection, analysis, and other activities not just in the summers but also in the academic years. Due to differing institutional constraints, the student at UCSB will be designated a "graduate student fellow," while the one at U. Miami will be a "graduate student researcher." (*Explanation*: Because the Mellon Foundation does not pay for the tuition and fees that at UCSB would normally be part of the compensation of a designated graduate student researcher [GSR], WE1S is asking for "fellowship" status for such students funded in part by a fellowship from the Mellon. This allows the university to trigger an internal provision that allows its Graduate Division and the home department of the graduate student to co-contribute tuition and fees on top of an external fellowship, thus bringing total compensation up to the same level as that of a normal GSR or teaching assistant. By contrast, U. Miami is able to designate the equivalent position a GSR because as a private university it has the flexibility to waive tuition and fees for the position.)

  (See *Appendix* for draft job descriptions for these positions.)

- *Graduate and undergraduate research assistants (hourly compensation) for summer, academic year, and technical programming work*. WE1S will also involve students as research assistants on an hourly wage basis for positions in the summers and academic years (and also others during either summer or academic year for technical programming). The purpose is not only to advance the project's research and technical goals but also to provide students with an opportunity to learn about digital humanities methods. The majority of graduate student RAs in the project timeline will be appointed during the summers *(see "Summer Research Camps" below in section II.e)*. Summers are when graduate students at the project's institutions are most in need of support and also when they are actually able to take RAships (because their allowed RA hours are constrained during the academic year by the terms of other support). There will also be some undergraduate research assistantships at CSUN designed to expose students to research activities and to encourage participation by diverse student populations. Some undergraduates from computer science or related departments may also be recruited to assisting in the project.

(See *Appendix* for draft job descriptions for the summer, academic year, and technical programming versions of these research assistant positions.)

- *Other faculty and staff researchers*. Faculty member Mauro Carassai at CSUN will participate in the project by helping to lead the work of research assistants at CSUN during the summer research camps. Additional faculty to be determined at CSUN will also participate in the summer research campus.
    - *Mauro Carassai, CSUN--*
      Mauro Carassai is Assistant Professor of Liberal Studies at California State University Northridge, where he teaches courses in digital humanities, literary theory, and American literature. He was a Brittain Postdoctoral Fellow at Georgia Institute of Technology in 2014-15 and a visiting Fulbright at Brown University in 2007-2008. His research combines literary theory, philosophy of language, and digital literatures within the larger frame of American literatures and American studies. His scholarly work has been published in journals such as *Culture Machine*, *LEA Almanac*, and *ADA – A Journal of Gender Media and Technology*. He co-edited a double issue for the *Digital Humanities Quarterly* titled "Futures of Digital Studies" and he is currently at work on a manuscript exploring problems and perspective in configuring an *Ordinary Digital Philosophy*. (For full c.v., see *Appendix*.)
- *Advisory Board*. Andrew Goldstone and Ted Underwood currently serve as advisory consultants to WE1S. The project plans to recruit a larger advisory board of about ten members total. This will include experts in digital humanities methods and also some whose work relates to WE1S's sociocultural themes (e.g., one or more scholars who are able to advise on the project's expanded aim of exploring the relation of underrepresented groups to the humanities). A possible candidate list for the advisory board is the following (many members of which have had previous contact with the WE1S team and are likely to agree to serve):
    - *Mark Algee-Hewitt*
      (Stanford U., English, Director of Stanford Literary Lab)
    - *David Bamman*
      (School of Information, UC Berkeley)
    - *Ryan Cordell*
      (Northeastern U., English, PI of NEH Digital Humanities Start-Up Grant, "Uncovering Reprinting Networks in Nineteenth-Century American Newspapers," core faculty of NULab for Texts, Maps, and Networks)

- *Amy Earhart*
  (Texas A & M, English, co-PI of TAMU's Initiative for Digital Humanities, Media, and Culture)
- *Gabrielle Foreman*
  (University of Delaware, Black American Studies & History, Director of Colored Conventions: Bringing Nineteenth-Century Black Organizing to Digital Life project)
- *Andrew Goldstone*
  (Rutgers U., English, creator of dfr-browser topic modeling interface)
- *Ryan Heuser*
  (Stanford U., English, Ph.D. candidate and former Associate Director for Research of the Stanford Literary Lab)
- *Laura Mandell*
  (Texas A & M, Director of TAMU's Initiative for Digital Humanities, Media, and Culture)
- *Trevor Muñoz*
  (University of Maryland, College Park, Assistant Dean for Digital Humanities Research, Associate Director of the Maryland Institute for Technology in the Humanities [MITH])
- *Lisa Nakamura*
  (U. Michigan, American Culture, Asian/Pacific Islander American Studies)
- *Safiya Umoja Noble*
  (UCLA, Information Studies)
- *Élika Ortega*
  (Northeastern University, Cultures, Societies & Global Studies; member of Global Outlook::Digital Humanities executive committee)
- *Andrew Piper*
  (McGill U., English, Director of .txtLAB, Editor of *Cultural Analytics*)
- *Benjamin M. Schmidt*
  (Northeastern U., History, core faculty of NULab for Texts, Maps, and Networks)
- *Richard Jean So*
  (U. Chicago, English; Center for the Study of Race, Politics, and Culture; PI of grant for "Computational Approaches to American Literature")
- *Robert B. Townsend*
  (Director, Washington Office of American Academy of Arts and Sciences; co-developer of Humanities Indicators; former Deputy Director of American Historical Association)
- *Ted Underwood*
  (U. Illinois, Urbana-Champagne, English, Information Sciences)

e. Activities and Timeline

During the three project years of its proposed Mellon grant, work on WE1S will be both iterative (repeated activities) and developmental (moving along a timeline through a series of goals). In the below description, "WE1S project team" refers to the combination of the project's PI, co-PIs, postdoctoral faculty fellows, research assistants, and Thomas Padilla in his role as Sustainability and Usability Advisor (with further specification of sub-groups organized by particular campus or task as needed). Each "project year" begins October 1 and ends September 30, coinciding with the academic calendar of an academic year and ensuing summer at UCSB.

1. Repeated Activities

Main categories of repeated activities in each project year are as follows:

- *Project meetings*: The WE1S project team will meet according to the pattern below (which proved to be effective in steering and coordinating the project during WE1S's pilot project phase):
    - The PI or co-PI(s) leading a particular campus or task group will meet with participating research assistants and postdoctoral faculty fellows to plan, coordinate, and discuss ongoing activity. The pace of these meetings will be dictated by the particular tasks and the phase of the project at the time. (For example, meetings during the summer research campus described below will be frequent.)
    - Approximately once per month WE1S will convene an all-project-team meeting (via a combination of face-to-face and remote conferencing). The purpose of these meetings is to share and coordinate work across the project's campuses and task groups.
- *Visits by PI and co-PI's to each other's campuses*: Once a year, the PI and co-PIs will travel to each other's campus to meet directly with a campus's research assistants and other participants.
- *Annual summer research camps*. Summer research camps at UCSB and CSUN will enlist student research assistants in the data-collection, technical, and interpretive work of the project. At each campus, RAs will be organized into two or three interdisciplinary teams (ideally a mix of humanities, social science, media-arts-and-technology, and/or computer-science students). Summer research camps will be approximately 6 weeks long. Activities will include: an orientation and training week at the beginning; interim weeks of combined group and individual work; and a final week devoted to presenting and discussing results with the whole WE1S team across the project's campuses (via remote conferencing). The RA teams will work on research related to data collection and the project's intellectual and technical methods. They will produce interim output for the project--for example, reports on current intellectual or technical issues faced by the project, white papers on specific topics, resources and recommendations for discussing the humanities in public, or digital exhibits or showcases. The final summer research

camps at the end of the project's third year will focus on the theme of "making it public"--i.e., brainstorming and prototyping ways to showcase the WE1S project and disseminate its outcomes to the project's three audiences of the public, humanities scholars and administrators, and digital humanities scholars. (See *Appendix* for the draft job description for summer RAs.)

At UCSB, WE1S will recruit an average of 12 graduate-student RAs each summer to be supervised by PI Alan Liu and co-PIs Jeremy Douglass and Lindsay Thomas. (Thomas will reside in Santa Barbara each summer for the research camps.) The goal of 12 RAs is stated as an average because the exact number each summer may vary depending on the availability of students able to work together during the same summer weeks. Students will be able to re-apply to participate in summer research campus in successive years, though WE1S will aim for a mix of returning and new students.[19] In project years 2 and 3, WE1S hopes to hire for the summer research camp at UCSB some RAs from other universities if their home institutions can subsidize the students' travel and lodging.[20] (Alternatively, some students from other institutions may be hired to work in the summer research camps through remote coordination and online meetings with the rest of the group.) The purpose of widening the RA pool to other institutions is to broaden the pool of talent/expertise/skills that the project can draw on, and also to broaden the project's impact by training other students in its ideas and methods.

At CSUN, the summer research campus will be structured much the same as at UCSB, but with the following differences adapted to the institution and its students. WE1S will recruit at CSUN up to 10 students at mixed levels (masters graduate students, plus a few undergraduates) to be supervised by CSUN faculty member Mauro Carassai. Up to two other CSUN faculty will likely participate in the summers. Students will also be able to re-apply to participate in summer research campus in successive years.

2. Timeline

WE1S's work will proceed as follows (with differences in each year's version of repeated activities noted):

i. Project Year 1 (2017-2018)

Year 1 will be an intensive development year in which WE1S will concentrate on these tasks:

[19] Some summer RAs at UCSB may also be hired on as academic year RAs. As usual in the case of hiring student RAs, exact pre-planning of staffing is difficult due to such factors as the availability of particular students, students' other funding support from their department/university units (which can determine by rule the maximum other work they can take on), and the timing of events in the students' educational career (e.g., their Ph.D. orals exams).

[20] Co-PI Douglass has successfully implemented this model previously for the annual research showcase of the UCSB English Department's Transcriptions Center (digital humanities and new media studies center), which he directs. Students from another California university visited UCSB with sponsorship from their institution.

- *Recruit and hire academic year lead RAs and other RAs.* PI Liu will recruit graduate-student research assistants at UCSB for the academic year (with work beginning as soon as possible after the Mellon grant's October 1 start date). He will also recruit the "graduate student fellow" at UCSB (the latter only for the winter and spring quarters this year because there will not be enough lead time after learning whether the project's Mellon proposal is approved to recruit a student for fall quarter). Co-PI Thomas will recruit and hire graduate students for the equivalent academic-year RAships and "graduate student researcher" positions at U. Miami (the latter only for the spring semester this year due to the timing of the Mellon grant start date). At UCSB, co-PI Douglass will also assist Liu in recruiting graduate and/or undergraduate RAs for programmer tasks.
- *Advertise for, interview, and recruit two "postdoctoral faculty fellows" at UCSB for project year 2 (with terms extending through project year 3)*. PI Liu will take the lead in conducting a national and international search for WE1S's two "postdoctoral faculty fellow" positions to start in project year 2. The search will conform to the calendar of the normal job-search season in humanities fields such as literature and history: advertisements in early fall, interviews with candidates in December or January (by Liu and the WE1S co-PIs via remote conferencing), and recruitment in early or late spring. *End-of-year result*: hiring of two postdoctoral fellows.
- *Recruit WE1S advisory board*. PI Liu will lead the recruitment of the WE1S advisory board. *End-of-year result*: establishment of the project advisory board.
- *Research and plan for expanding the WE1S main corpus and sub-corpora:* Beginning as soon as possible after the onset of the Mellon grant, PI Liu and Co-PI Thomas will lead the WE1S project team (assisted by the RAs each supervises at their campuses) in investigating sources of new documents for collection in the project's dataset (according to criterial detailed above in *section II.b.1*). This investigation will involve an iterative process of defining the main corpus and various sub-corpora, considering potential sources of materials, and devising any additional or variant methods needed to harvest data from these sources beyond those WE1S has already developed. To facilitate research into how the humanities are viewed by, or in relation to, different racial, ethnic, gender, immigrant, and age groups, Liu and Thomas will also continue efforts they have begun to consult with scholars and university administrators whose work relates to such groups. Additionally, they will consult with the WE1S advisory board at the project's conference meeting on these issues as well as on directions for the other possible sub-corpora (from historical, government, scholarly, and other documentary sources as mentioned in *section II.b.1*). *End-of-year result*: the goal by the midpoint of project year 1 is a prioritized (if still evolving) list of source material to stage for collection during the rest of project year 1 and in project years 2 and 3. The collection list will be stabilized by the end of year 1.
- *Evolve the WE1S "random" sample corpus*. Co-PI Thomas will lead RAs at U. Miami in evolving the "random" comparison corpus they have already started and conducting comparison research with it (see under the task labeled "Experiment with

complementary methods of analysis" task below). *End-of-year result*: A "random" sample corpus and comparison experiments using it.

- *Experiment with alternative search methods for identifying materials to collect*. Currently, WE1S identifies materials to include in its corpus by searching for mentions of "humanities," "liberal arts," and (in publications in the United Kingdom or Commonwealth nations) "the arts." In coordination with the corpus expansion work led by Liu and Thomas, co-PI Douglass will early in project year 1 take the lead (assisted by programmer RAs at UCSB) in experimenting with other kinds of searching--for example, identifying articles relevant to the humanities by looking for collocated phrases in near-proximity and developing ways to search sources based on these more nuanced signals. (Such experiments may help validate WE1S's existing search criteria even if their methods are not ultimately adopted either because they are technically impractical under the constraints of source databases or because they do not locate enough additional or different material to justify the extra complexity.) *End-of-year result*: decision on whether or not to adopt alternative or additional searching methods.
- *Experiment with complementary methods of analysis*. PI Liu and co-PI Douglass will lead a group of RAs at UCSB familiar with word2vec ("word embedding") and other text-analytical methods in initial word-embedding explorations of samples from the WE1S corpus. These experiments will help determine if analytical methods other than topic modeling can be useful and also technically feasible in complementing the primary research methods of the project. Co-PI Thomas will meanwhile lead RAs at U. Miami in experiments with other kinds of analytical methods--e.g., training classification algorithms--to explore linguistic features that differentiate materials discussing the humanities from other materials in the project's "random" sub-corpus. *End-of-year result*: assessment (through study and discussion) of experimental results to determine whether to plan for further development of these analytical methods in project year 2.
- *Improve technical methods of collecting primary materials for analysis ("ingest")*. Co-PI Douglass will lead the project team (with the assistance of RAs at UCSB) in improving technical methods for ingesting textual materials for topic modeling. As described in *section I.f, Technical Methods*, WE1S has developed during its pilot project a technical environment for its research. The slowest part of its technical workflow at present is the initial one. The collection of materials and their conversion into analytical data prepared for topic modeling has involved much manual work extending from initial collection to preliminary pre-processing of results into data. While not all manual work can be avoided, since the terms and conditions of some sources prohibit algorithmic searching and downloading, Douglass will develop improvements to WE1S's ingest methods. PI Liu and co-PI Thomas will coordinate with Douglass on materials to submit to new ingest methods for testing purposes. Co-PI Kleinman will coordinate with Douglass to ensure that materials are ingested in a form consistent with the WE1S manifest framework. *End-of-year result*: an improved ingest system.
- *Purchase hardware and software (plus possible cloud-based services) to advance the project*. Co-PI Douglass will lead in assessing the existing computer hardware and

software/cloud-based platform for WE1S's technical environment and deciding what needs to be purchased and installed to advance the project's work. (For the relatively small amount of upgrades WE1S expects to need, see under *Budget Narrative.) End-of-year result*: a stable hardware and software/cloud-based platform for the project's work in the next years.

- *Begin collecting materials for the expanded WE1S corpus and sub-corpora*. After the preliminary research and technical development tasks described above, WE1S will begin collecting new materials at scale. The collection task will be supervised by the PI and co-PIs, and will extend from the academic year into WE1S's first summer research camp. The amount of new material to be collected in project year 1 (measured in numbers of publication sources) is hard to predict, since it depends on the results of research into what to collect and when improvements to the technical collection process are ready to be implemented. WE1S's goal is to be able to collect at scale by approximately midway through project year 1 (e.g., beginning of spring quarter of the UCSB academic year 2017-18). *End-of-year result*: progress on collecting materials for WE1S's corpus expansion.
- *Extend the WE1S interface for dynamic exploration of topic models.* WE1S has implemented Andrew Goldstone’s dfr-browser as a method of visualizing and exploring its topic models. However, because dfr-browser was originally customized for topic models created only from documents in JSTOR, it does not work out of the box with the newspaper articles and other sources (and their metadata) used by WE1S. With the assistance of Andrew Goldstone, WE1S has adapted dfr-browser for its materials in a usable but incomplete way. Co-PIs Thomas and Kleinman will work to complete the adaptation of dfr-browser to WE1S by customizing the dfr-browser code to use metadata provided in the WE1S manifest framework and thus to provide fully dynamic access to WE1S data. If it proves feasible, WE1S will also see if additional kinds of visualizations (e.g., of topic words or topic clusters) such as those generated by the Lexos text analysis site can augment dfr-browser in facilitating the exploration of topic models. *End-of-year result*: fuller adaptation of dfr-browser for WE1S's materials (and possible additional visualization methods).
- *Integrate the WE1S Workflow Management System and Virtual Workspace System*. During its pilot project, WE1S developed its Workflow Management System and Virtual Workspace System separately in response to different project needs. The Workflow Management System describes the collections and processing of materials, while the Virtual Workspace System implements the actual workflow. During project year 1, co-PIs Douglass and Kleinman will integrate these systems into a single workflow environment. The goal is to allow workflows described in the Workflow Management System to be run in the Virtual Workspace System and, reciprocally, jobs run in the latter to generate manifest descriptions of the sources and processes to be recorded In the former. *End-of-year result*: integration of the WE1S Workflow Management System and Virtual Workspace System.

- *Evolve the WE1S interpretation protocol for topic models*. PI Liu will lead the WE1S project team in using samples of the project's corpus of materials to generate initial topic models that can be used to rehearse procedures for the assessment and interpretation of topic models. These rehearsals (conducted through study and discussion of topic models, as well as through reflection on the process of study and discussion itself) will be the basis for evolving a topic-model interpretation protocol that can be declared and shared as one of the project's outcomes. *End-of-year result*: draft version of interpretation protocol.
- *Begin planning and modeling a public-facing Web site.* Working with Thomas Padilla, WE1S's Sustainability and Usability Advisor (and assisted by RAs), co-PIs Douglass and Kleinman will lead initial planning for a public project Web site that can present topic models, analyses, and other outcomes. Such planning will involve considering how best to utilize WE1S's internal project workflow management systems (its Workflow Management System and Virtual Workspace System) to generate content for a public interface. It will also involve identifying the features most useful for WE1s's three user audiences (the public, humanities scholars and administrators, and digital humanists) as well as issues of sustainability. *End-of-year result*: initial concepts and plans that set priorities for year 2 development.
- *Convene advisory board meeting at UCSB in spring quarter*. Assisted by RAs, PI Liu will organize and lead the conference meeting of the WE1S advisory board in spring 2018 to discuss the aims, methods, materials, and goals of the project to date. (The meeting is delayed until late in the first project year due to the lead time needed to recruit the advisory board and organize travel, lodging, and other logistics.) The conference meeting will be structured around an initial series of short presentations of WE1S project work, followed by extended discussion with the board. The outcome of the conference meeting will be specific recommendations by the board for revising or improving WE1S's development plans for years 2 and 3.
- *Plan summer research camps and recruit RAs for them.* At UCSB, PI Liu will lead in detailed planning of the schedule and activities of the first year's summer research camp as well as in recruitment of RAs for them. At CSUN, Co-PI Kleinman will lead in the equivalent planning and recruitment work for the summer research camp on his campus. He will also meet with CSUN campus representatives of the Mellon-funded HSI Pathways to the Professoriate program to recruit student participants from that program. Additionally, he and Professor Mauro Carassai will recruit faculty participants for the CSUN summer research camp. The structure and activities of these research camps are described above (under repeated activities). The specific emphasis of the first year's research camps will be to accelerate and advance the activities of project year 1 described above--specifically, by asking RAs to collect materials for the WE1S corpus, engage in shared interpretation of topic models, help evaluate the effectiveness of alternative analytical methods, and discuss the social and public implications of preliminary findings so as to anticipate what the eventual public-facing Web site should emphasize. *End-of-year results*: the outcome of the summer camps will be additional materials collected for the WE1S corpus along with student research reports,

presentations, and/or projects (such as initial analyses or visualizations of interpretive results or humanities advocacy recommendations).

- *Recruit academic-year research assistants for the next project year*. By the end of summer, PI Liu will recruit new or continuing RAs and a new "graduate student fellow" for the ensuing academic year at UCSB. Co-PI Thomas will do the equivalent at U. Miami.

ii. Project Year 2 (2018-2019)

In Year 2, WE1S will concentrate on advancing the development tasks started in the project's first year and positioning the project for completion in the final year:

- *If needed: advertise for, interview, and recruit new "postdoctoral faculty fellows" at UCSB for project year 3*. While the postdoctoral faculty fellows hired during year 1 will have the option of continuing for a second year, it is sometimes the case that postdocs find tenure-track jobs or for other reasons choose not to continue. If this is the case, PI Liu will recruit new postdocs for project year 3. The advertising, interviewing, and recruitment schedule may need to begin later than fall, depending on when WE1S learns whether its postdocs will be continuing or taking another job.
- *Continue expanding the WE1S main corpus and sub-corpora*. PI Liu and co-PI Thomas will lead this activity with the assistance of RAs on their campuses. Goals for this year include continuing to collect data from the sources identified in the previous year as well as evolving or revising the target list of publications and other materials based iteratively on topic modeling and studying the materials already collected, and on assessing the overall pace of collection work (which may lead to adjusting the size of the target list of materials). *End-of-year result*: a nearly complete set of project data as well as identification of any materials that can still be realistically ingested during the rest of the project timeline.
- *Topic model and study the extant WE1S corpus of materials*. The PI and all co-PIs will lead the project team at their campuses (postdocs and RAs) in iterative topic modeling of parts of WE1S's corpus of materials (as it exists at the time) and in the interpretation of results according to the project's evolving topic-model interpretation protocol. While still interim in status, these topic models and analyses will help the project team assess how well WE1S's research methods and technical implementation are performing in addressing the research hypotheses mentioned in *section I.b* above. The interim results may lead to revising the target list of materials to collect for the project's main corpus and sub-corpora, or to revising research and technical methods. The results will also aid in development work this year on planning for the WE1S's public-facing Web site and the project outcomes it will present. *End-of-year result*: a set of interim topic models and analyses.
- *If feasible: experiment with collecting and topic modeling a small sample of Spanish-language materials*. With the assistance of RAs (and in consultation with colleagues familiar with Spanish-language journalistic materials and Chicana/o studies), PI Liu and co-PI Thomas (who is fluent in Spanish) will take the lead in attempting to collect a

relatively small sub-corpus of Spanish-language newspaper articles related to the humanities. While it is not possible to include such material in the same topic model with English-language materials, a separate topic model of the Spanish sample will allow WE1S to experiment with comparison work (e.g., through comparing how approximately similar topics are weighted or clustered in the two topic models).

- *Advance the exploration and implementation of alternative search methods and complementary analysis methods*. If at the end of year 1 the exploration of alternative search methods results in a decision to commit to implementation, then co-PI Douglass (assisted by RAs) will lead the task of integrating such methods in the project's technical workflow. In addition, if at the end of year 1 the exploration of complementary text-analysis methods such as word embedding and classification algorithms led to a decision to continue exploration, WE1S will in year two assess the results to date to see if their value in confirming or augmenting the results of topic modeling justify implementation. If so, Douglass will be assisted by RAs in building these methods into the project technical workflow. *End-of-year result*: implementation of these additional methods in the WE1S workflow (contingent on a decision to commit to them).
- *Continue to improve the operation and integration of the WE1S Workflow Management System and Virtual Workspace System; and begin developing the WE1S repository system(s) for sustainability.* Co-PIs Douglass and Kleinman will modify the WE1S Workflow Management System and Virtual Workspace System in response to any needs identified by the project during year 2. In addition, they will work with Thomas Padilla to begin planning for the way these systems and the research materials they generate (e.g., topic models, workflow manifests, Docker containers, research analyses and reports) will be placed in a repository system for institutional deposit as well as disseminated on GitHub (as described in *section II.h, Sustainability*). *End-of-year result*: improvements to the Workflow Management System and Virtual Workspace System, and a plan for implementing the project's preservation and dissemination repository systems in the final project year.
- *Finalize the WE1S interface for dynamic exploration of topic models*. Based on any further needs identified by the project team (and with additional consultation from the project advisory board), co-PIs Thomas, Kleinman, and Douglass will make further customizations of the dfr-browser interface and complete the integration of any additional exploration/visualization tools. *End-of-year results*: WE1S's final interface for exploring topic models.
- *Advance the WE1S interpretation protocol for topic models*. Based on the project team's experience working with interim topic models during the first two years (and after further consultation with the project advisory board), WE1S will finalize a version 1.0 protocol for the interpretation of topic models (written up both as a document and as a "manifest" in the project's manifest framework). PI Liu will lead this activity, in which all project participants will contribute. *End-of-year result*: version 1.0 topic modeling interpretation protocol.

- *Prototype WE1S's public-facing Web site*. Extending the work on this task from the first year, co-PIs Douglass and Kleinman will lead the development of a prototype public-facing Web site that draws some of its content dynamically from the WE1S Workflow Management System and Virtual Workspace System and also allows for the addition of written analyses, reports, examples, and recommendations. In consultation with its advisory board, WE1S will commit to a design and feature set by the end of year 2. *End-of-year result*: A prototype public Web site.
- *Plan summer research camps and recruit RAs for them*. As in year 1, PI Liu will lead in planning the schedule and activities of the second year's summer research camp at UCSB as well as in recruiting RAs for them. At CSUN, Co-PI Kleinman and Professor Mauro Carassai will do the equivalent for the summer research camp on their campus. The structure and activities of these research camps are detailed above under repeated activities. The specific emphasis of the research camps in project year 2 will be to continue collection work at a rapid pace while also engaging in interpretation of topic models and of any other analytical results (e.g., word embedding analyses). Such interpretation conducted through study and discussion within each camp and collaboratively between camps (via remote conferencing) will be aimed at producing mock-ups of final outcomes--e.g., examples of topics, conclusions, reports, and advocacy recommendations to showcase on the project pubic Web site. As mentioned earlier, the summer research camp at UCSB may this year include some RAs from other universities outside the project.
- *Conduct "triage exercise*." At the end of the summer in year 2, the WE1S PI, co-PIs, postdocs, and lead RAs will meet to review each continuing task in the project to identify any problem areas that need to be tied off to allow WE1S to bring its research to a conclusion by the end of year 3. For example, the project team will ask:
    - Are there publications still on the target list for collection that the project cannot feasibly get to during the remaining timeline?
    - Are there planned features in the WE1S Workflow Management System or Virtual Workspace System that cannot be implemented during the timeline?
    - Can word embedding or other additional analytical methods procedures be implemented at scale and used for analysis during the timeline? (If not, can WE1S create smaller-scale implementations with samples of material instead?)
    - What features and functions desired for the public-facing Web site must be let go to allow the site to be finished in time?
- *Recruit academic-year research assistants for the next project year*. By the end of summer, PI Liu will again recruit new or continuing RAs and a new "graduate student fellow" for the ensuing academic year at UCSB. Co-PI Thomas will do the equivalent at U. Miami.

iii. Project Year 3 (2019-2020)

The final year of the WE1S timeline will be devoted to finishing collection and development tasks, creating final topic models and analyses, and disseminating outcomes to the public and scholars. Tasks will include:

- *Finish collection work for the WE1S main corpus and sub-corpora, and create a "scoping statement" for the collection*. Activities related to collecting and ingesting materials as datasets will be completed near the beginning of year 3 so that WE1S can concentrate on analysis and dissemination work. PI Liu and co-PI Thomas, with the assistance of RAs at their campuses and also in consultation with other co-PIs and postdocs, will take the lead in writing a scoping statement describing the nature, selection criteria, and organization of the project's gathered materials (with their associated manifests providing metadata on provenance and workflow) so that WE1S's public, humanities scholar and administrator, and digital humanities audiences will be able to understand what was gathered for study. *End-of-year result*: a finished corpus and sub-corpora, with scoping statement.
- *Create a set of topic models that will be the basis of disseminated analyses, examples, and reports*. Led by PI Liu, all the project co-PIs, postdocs, and RAs will participate in creating and interpreting a set of topic models of the WE1S materials. If in year 2 WE1S committed to additional methods of text analysis such as word embedding, models based on those methods will also be created and interpreted. Specifically, the project team will conduct a series of analysis meetings during the year (extending into the final summer research camps) that concentrate on understanding different findings from the models. In the case of topic models, for instance, questions to be examined include:
    - What are main and outlier themes in public discourse related to the humanities?
    - How do the pattern of those themes, their relative weights, and their distribution across publications and articles (e.g., in articles that seem on their face to be about some other area such as politics or science) help us understand the general configuration of "the humanities" in public discourse?
    - Can temporal trends be identified?
    - Can differences in the distribution (and weights) of topics across publications--e.g., publications of one kind or another, from one nation versus another, or addressed to one social group versus the mainstream--be understood as socially or culturally significant?

  End-of-year result: a set of publicly showable topic models, plus interpretive results to be disseminated through analyses, reports, and examples.
- *Prepare analyses, reports, and examples based on WE1S's research*. PI Liu will organize the project team into groups and/or individuals responsible for producing by the end of the project year a set of analyses, reports, and examples or exhibits (e.g., curated examples or features of topic models with links to exemplary articles in their original publications). These will be designed to address the project's different audiences. For

example, one group of co-authors may be assigned the task of creating such outcomes for the general public; another for humanities scholars and administrators; and yet another for digital humanists. Depending on the evolution of the Hypothes.is organization's Web annotation system (see under *section II.b.2* above), WE1S may use its standing as one of Hypothes.is's "public groups" to create a curated annotation and highlight "layer" over newspaper or other publication Web sites. This will demonstrate *in situ* WE1S's findings. (Such annotations can include links to analyses published on WE1S's own public Web site.) *End-of-year result*: a suite of analyses, reports, and examples.

- *Prepare recommendations and resources for humanities advocacy based on WE1S's research.* PI Liu will lead the project co-PIs, postdocs, and RAs in producing a set of recommendations and resources for humanities advocacy. Recommendations may take the form of executive summaries addressed to different sectors of the public (e.g., journalists, politicians, business leaders, parents, students) and best-practices advice (e.g., avoiding untrue or overused themes in public discussion of the humanities, and drawing connections between the humanities and themes of interest to the public). Resources might take the form of "kits" of themes, examples, and evidence for journalists or scholars to draw on in discussing the humanities; or students to draw on as they consider choosing a major and discussing it with their parents. *End-of-year result*: a suite of recommendations and resources.

- *Prepare an overview description and rationale statement for the WE1S project addressed to its overlapping audiences of the public, humanities scholars and administrators, and digital humanists.* PI Liu will lead the project team in writing an overview description and rationale statement that frames the WE1S project for public view. This statement will include context about other work and projects related to the state of the humanities (of the sort instanced in *section I.b, Humanities Context* above). *End-of-year result*: A description and rationale statement ready for presentation on the WE1S public-facing Web site.

- *Complete the WE1S public-facing Web site*. Co-PIs Douglass and Kleinman will complete work on the WE1S public-facing Web site, which will draw on such features of the project’s Workflow Management System and Virtual Workspace System as the ability to visualize topic models for exploration and to provide links back to original documents (where not restricted). The Web site will also present the description and rationale statement for the project; the scoping statement created by Liu and Thomas; the analyses, reports, and examples produced by the project team; and WE1S's recommendations and resources for humanities advocacy. In addition, the Web site will have an API or other means of exporting the project’s data (only analytical, provenance, and other data that can be shared without restriction) for research by others. *End-of-year results*: publication of the WE1S Web site.

- *Plan and implement a publicity and social media campaign to disseminate WE1S outcomes*. With the assistance of RAs and consultation of the rest of the project team, PI Liu will lead the planning and implementation of a publicity and social media campaign for the project. This will likely involve the writing of news releases that can be circulated to media, state humanities councils, humanities centers, and individual scholars and administrators; involving the principals and other project members in giving interviews to local and other media; and preparing a series of social media postings timed for the beginning of the new 2020-2021 academic year.

- *Disseminate scholarly output*. Individually and as co-authors, the WE1S project team will begin preparing materials that can be presented as talks at academic conferences, articles in journals, and other traditional forms of humanities scholarship. All project principals will participate in this activity, and will encourage postdocs and graduate students to do the same as individual authors or co-authors. Target conferences include both those in the digital humanities and in general humanities fields, such as the conferences, respectively, of the Alliance of Digital Humanities Organizations and the Modern Language Association. Target publications include journals for digital humanities audiences, such as *Digital Humanities Quarterly*, *Digital Scholarship in the Humanities*, and *Cultural Analytics*; and also journals for broader fields of humanities scholarship, such as *PMLA*, *New Literary History*, and *History of the Humanities*.
- *Commit project materials for preservation and dissemination*. Led by co-PIs Douglass and Kleinman, and with the aid of Thomas Padilla, WE1S will finalize the WE1S repository system for institutional deposit as well as for dissemination through GitHub (as described in *section II.h, Sustainability*). *End-of-year results:* all WE1S outputs that may be made public, including technical environment and metadata, topic models, analyses, reports, and recommendations will be deposited in institutional and GitHub repositories.
- *Plan final summer research camps and recruit RAs for them*. As in years 1 and 2, PI Liu will lead in planning the schedule and activities of the final year's summer research camp at UCSB as well as in recruiting RAs for them. At CSUN, Co-PI Kleinman and Professor Mauro Carassai will do the equivalent for the summer research camp on their campus. The structure and activities of these research camps are detailed above under repeated activities. The specific emphasis of the research camps in the final project year will be to contribute to WE1S's dissemination phase by concentrating on the tasks outlined above of preparing analyses, reports, and examples; preparing recommendations and resources for humanities advocacy; and assisting in the creation of materials for a publicity and social media campaign. Interdisciplinary teams of RAs at the two camps this summer will be organized to work on these tasks for specific audiences. For example, one or more teams will help create materials addressed to the general public; another team(s) will do so for humanities scholars and administrators; and yet another team(s) will do so for digital humanists. As in year 2, the final summer research camp at UCSB may include some RAs from other universities outside the project.

f. Expected Outcomes and Benefits

Expected Outcomes

Outcomes at the close of the WE1S project timeline will include a public-facing Web site presenting:

1. An overview description and rationale statement for the project.
2. A scoping statement of the materials collected and studied.
3. Topic models presented in a dynamic, interactive interface (based on dfr-browser) designed to encourage users to explore topics and read exemplary source articles (linked or cited in their original locations). If the project utilizes word embedding or other additional analysis methods (described as contingencies in the project timeline and *section I.e, Research Methods*), the models generated by these approaches will also be featured.
4. Analyses and reports on what the project's research brings to view about public discourse on the humanities. These materials will include (or link to) examples in such forms as: galleries of quotations, sample newspaper articles (cited or linked in their original locations), lists of evidence or anecdote types often used in discussion of the humanities, and a Hypothes.is "public group" layer of annotations and highlights over selected media articles.
5. Recommendations for humanities advocacy in the form, for example, of executive summaries addressed to different sectors of the public (journalists, politicians, business leaders, parents, students) and best-practices advice (e.g., avoiding untrue or overused themes in public discussion of the humanities, and drawing connections between the humanities and themes of interest to the public). WE1S will also tap its project team (including student RAs at its final summer research camp) for other creative ideas, such as "rewriting" a media story about the humanities.
6. Resource "kits" of themes, examples, and evidence for journalists or scholars and administrators to draw on in discussing the humanities; or students to draw on as they consider choosing a major and discussing it with parents. WE1S will also draw on its project team (including student RAs at its final summer research camp) for other creative ideas, such as producing infographics, timelines, and storymaps.

(Though it would be ideal for WE1S to assess systematically the effectiveness and audience appropriateness of its recommendations, summaries, and resource "kits" *[items 5 and 6 above]*, such appraisal falls outside the scope of the currently proposed project. In strategizing what journalistic materials to add to its corpus, WE1S has consulted recent research on "media impact" [e.g., Schiffrin and Zuckerman] that may help in such assessment in the future. Potentially, a next stage of the project could use the criteria and methods explored in media impact research to assess WE1S's own public outputs as well as other humanities advocacy.)

In addition, expected outcomes include the deposit of the WE1S technical systems (its manifest framework, Workflow Management System, Virtual Workspace System, and topic modeling interpretation protocol) in an institutional repository as well as a GitHub repository. These deposits will include manifests documenting the work of the WE1S project (e.g., how a topic model was produced) but no primary materials owned by other parties.

Expected Benefits

For its overlapping audiences of the public, humanities scholars and administrators, and digital humanists, the benefit that frames all subsidiary ones will be to use research-based knowledge to advance a more expansive notion of humanities advocacy--one that makes its beneficiary not just humanities disciplines, scholars, and students but the larger public. By exploring its research hypotheses *(see section I.b)* on such issues as the relation between focalized and "general life" understandings of the humanities, and on the way different nations and social groups view, or are viewed in relation to, the humanities, WE1S will be able through its recommendations to depict how being a good educator (whatever the field), journalist, business person, politician, technologist, or parent involves engagement with the humanities at some level. In this expanded sense, humanities advocacy is about using knowledge about, and gained through the humanities (in this case specifically the digital humanities), to advocate for being a good educator, journalist, business person, politician, technologist, parent, or child *as such*.

Specific, concrete benefits for WE1S's audiences include:

- *For various sectors and professions among the public*, WE1S will provide a richer stock of themes, narratives, examples, and evidence types that can be drawn upon in discussing humanities-related issues, whether at the policy level (e.g., how society should apportion investment in STEM versus humanities education) or at the individual or social level (e.g., how parents and students talk to each other about what they want to do in life). WE1S will also help widen the social and cultural diversity of public discourse on the humanities, bringing into consideration not just the generic "student" or "major" referred to in many media stories but also students whose specific racial, ethnic, gender, immigrant, and generational backgrounds positions them differently in the field of such discussions.

  Equally beneficial will be the context in which WE1S frames this richer, more diverse mix of views on the humanities through its description and rationale statement, scoping statement, and analyses and recommendations. Simply creating a cognitive map situating discussions about the humanities in a more wide-ranging discursive field touching on issues important to society is valuable to offset tunnel-vision understandings about a future led by just a few technical or business-oriented professions.

  A concrete example of a benefit in the public sphere might be instanced in the following scenarios: a journalist is assigned to write this year's story about "the decline in humanities majors"; or a politician is pressed by her constituents to protect funding for humanities programs at the federal or local level. In such cases, WE1S will provide talking points to draw on, orientation about overused versus less frequently mentioned themes, insight about how the humanities have been discussed in relation to particular groups, and links to exemplary articles and other material.

- *For humanities scholars and administrators*, WE1S's outcomes will not only facilitate their own participation in humanities advocacy (by widening and enriching the discourse of such advocacy in the ways described above) but also augment specific research, program-building and administrative, and public outreach missions.

  In regard to research, the project's methods and tools will serve as a paradigm (and can be used "as is" or in adapted form) for researching the way other complex ideas that are like "the humanities" in having both narrow/sharply defined and broad/fuzzy senses behave in public discourse. Examples might include such concepts as "neoliberalism," "climate," "globalism," "science," or "culture." In addition, WE1S will provide research material for scholars working specifically on areas such as university studies or the history of the humanities.

  In regard to program-building and administration, WE1S will be able to assist in such activities as designing general education curricula, shaping agendas for humanities centers, or presenting a matriculation or commencement speech--all of which can benefit from WE1S's research-backed identification of themes and the relations between themes.

  And in regard to public outreach for the humanities, WE1S will provide an extended, enrichened range of themes, arguments, examples, and other material to draw on in framing advocacy efforts on behalf of funding for the humanities, bringing new students into humanities majors, and showing the connection of the humanities to other educational fields and to other areas of social concern. Especially valuable is the fact that WE1S's research materials and methods will allow for comparative methods of advocating the humanities that would not otherwise be easily available (except in traditional universalizing terms). For example, the scope of WE1S's studies will allow administrators and scholars (and university communications or public relations officers) to speak of the cross-regional or -national significance of the humanities, of the importance of the humanities across social groups, and of the participation of the humanities *alongside* the sciences and other fields in topics of great contemporary concern.

- *For the digital humanities research community*, WE1S will provide a paradigm of open, shareable, and reproducible research adapted for the kinds of provenance tracking, analysis workflows, and self-reflective attention to interpretive method characteristic of humanities-oriented "cultural analytics." The WE1S manifest framework, Workflow Management System, Virtual Workspace System, and topic model interpretation protocol will be disseminated so that they can be used or adapted by other projects. More generally, however, their value will lie in advancing a model for how data-intensive digital humanities research will need to be conducted in the future to meet currently emerging standards of openness, shareability, and reproducibility (e.g., of the kind now required for authors of articles in the *Cultural Analytics* journal; *see under I.c* above). WE1S imagines a scholarly future for the digital humanities in which critical rigor means not just showing and explaining results, but showing the underlying dataset (or only the derived analytical data if the primary data is proprietary) and also the data workflow. Through its summer research camps and other RA-related activities, WE1S will also have the benefit of training members of a new generation of graduate students in such scholarship.

g. Intellectual property

WE1S will produce two kinds of output that will be made available for use by others in as open a way as possible (constrained only by the underlying intellectual property restrictions of the project's source material). The outputs are:

- *A public-facing Web site presenting the project's research in the form of topic models, metadata, visualizations, the project's own authored documents (analyses, reports, recommendations, descriptions of resources), and other explanatory material.* These materials will be under a Creative Commons BY-SA 4.0 license (allowing for the reproduction and adaptation of material, subject to providing attribution and "sharealike" status). Topic models will be presented and visualized using interfaces such as dfr-browser that are open for use without needing permissions. Of course, newspaper articles and other original source material for WE1S's topic models under copyright by others will not be reproduced on the WE1S Web site. However, they will be cited or linked to in their original locations as needed. In addition, WE1S will use the open annotation Hypothes.is platform to layer comments over newspaper and other articles in their original online locations.
- *The technical research environment created by WE1S.* The technical research environment described in *section I.f.2-5* above (under *Technical Methods*) will be deposited in an institutional repository and also disseminated through a Github repository for open use by others. The WE1S Workflow Management System and Virtual Workspace System, along with their components (e.g., the WE1S manifest schema), will be under the MIT License.[21] Underlying components / platforms incorporated in these systems are themselves open source (e.g., Jupyter notebooks, MongoDB, Docker containers, dfr-browser). For example, dfr-browser, which WE1S incorporates as a main means of presenting topic models, is under the MIT License.

WE1S has researched the licensing and intellectual property constraints of such databases as LexisNexis from which it will draw materials--especially as regards constraints on automatic harvesting and the storing of downloaded materials. WE1S will implement a non-consumptive research workflow in which use-protected materials are not held by the project and never made available to workers in the research environment. In the workflow, original article content is downloaded to a secure machine with no direct user access. Each original article is deleted immediately after being transformed into a non-consumptive analytical data file (an “extracted metadata” or "bag of words" representation of a document consisting of text that has been alphabetized or, the equivalent, lists of unique words with frequency metadata). A property of

[21] MIT License: "Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions: The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software."

these analytical data files is that they cannot be used to reconstruct the original article; they can thus be stored without contravening copyright or licensing rules. Analytical data files are then posted to another computer, and can be mounted into the WE1S research environment for use in topic modeling workflows. Equivalent measures for storing only analytical data files that cannot be used to reconstruct original documents will be implemented if WE1S also pursues a word embedding (word2vec) or other text-analysis approach. In the case of word embedding, for example, non-consumptive lists of ngrams (consecutive sequences of words of various word lengths) will be stored in a text file for each article. In addition, ngrams will be stored in a “noised corpus” format, with a small number of fictitious ngrams added at random to prevent even the theoretical reconstruction of the original article (see Gallé and Tealdi). These non-consumptive use safeguards are designed to satisfy fair use doctrine for conducting research on large document collections as affirmed in *Authors Guild Inc., et al. v. Google, Inc.* and *Authors Guild, Inc. v. HathiTrust*.

h. Sustainability

To ensure the sustainability of WE1S data, WE1S will work with UCSB's Data Curation Program, which was launched at the UCSB Library in 2016 and is staffed with curators with expertise in data curation and preservation. Thomas Padilla, the Humanities Data Curator in the program, serves as WE1S's Sustainability and Usability Advisor (acting as an embedded member of the project team *[see section II.d, Staffing]*). Under Padilla's guidance, WE1S will ensure the long-term sustainability and reuse of project results according to the following plan:

Scope of Project Outcomes and Data: As described in *sections I.d and II.f*, WE1S's results will include analytical materials and data (e.g., topic models) intended for the public and also for humanities scholars (not including materials and data under the copyright or licensing limitations of others). These will be presented on a public-facing Web site; and will consist of structured and unstructured text documents (.txt, .pdf, .json, .html), scripts (.py, .R), Docker files for implementing open source Docker images and containers, a mixture of qualitative and quantitative data (.csv, .txt), and a MongoDB database that tracks data provenance as it moves throughout the WE1S system. Due to copyright and licensing restrictions, primary data from proprietary databases (e.g., newspaper articles) are excluded from long term preservation and access.

Preservation & Access of Project Materials and Data:

- Near-term preservation of working materials and data: For the duration of the Mellon grant and extending further for at least an additional seven years for a total of ten years (as committed to by UCSB's English Department), WE1S's analytical data, Web-based tools, and public Web site will be sustained on servers and backup storage (with nightly backups) controlled by the UCSB English Department. The local staging of development operations is important for agile, rapid work. The UCSB English Department, which has long-standing strength in the digital humanities, is the only humanities department at UCSB to run its own servers and to have technical support from a dedicated Computer and Network Technologist staff person.

- Long-term preservation of public materials and data: For long-term preservation and access, WE1S will regularly deposit materials and data meant for public consumption in the Zenodo repository system. A Zenodo integration for Github will also enable automated deposit of project code as it is developed within Github's version control repository environment. Zenodo is the European Commission's data repository for "open science." It is one of the world's most advanced, open, and customizable data-repository platforms. It accepts deposits from all nations and disciplines in a variety of formats. It assigns DOIs for deposits and commits to indefinite preservation. It also amplifies research impact through automated citation and discoverability methods.

  WE1S may consider migrating project materials and data to Humanities Commons, whose Fedora-based digital repository system will soon be moving to Fedora 4. Humanities Commons is a social network and repository system supported by the Modern Language Association. Fedora is an open source repository system that research libraries on an international level use to manage and disseminate large and complex digital collections of historic and cultural materials as well as scientific data. According to Kathleen Fitzpatrick of the MLA and the Humanities Commons staff (which whom WE1S has consulted), a future version of Humanities Commons may allow for plugins and other technical features that better accommodate some of the more specialized data like its assortment of Docker files.

  For very-long-term and end-of-project-life stages, WE1S will explore solutions for graceful degradation. WE1S's PI has had experience "flattening" complex, dynamic digital humanities projects into static online forms for long-term preservation. (He is collaborating with expert researchers in digital curation on preserving his Voice of the Shuttle Web site [vintage 1994] in ways that hybridize "flattening" with auto-generated links for extinct sites from the Internet Archive.) WE1S will explore very-long-term sustainability solutions as they develop.

i. Risks and Mitigation

The risks that WE1S foresees are characteristic of many digital humanities projects, though specifics differ. The main risks WE1S has identified are the following:

- *Risk due to loss or change of project faculty*.

  While changes in key project personnel can impact any project due to separations and other reasons, such risk takes special forms for digital humanities projects whose personnel includes junior faculty. Such faculty will be coming up for tenure based in part on non-traditional digital research.

  This risk is mitigated for WE1S because the two junior faculty on its core team (two of its three co-PIs) were specifically hired by their institutions into digital humanities/new media positions. They were given the understanding (in one instance contractually, in the other via precedents worked out in previous promotion cases) that digital-humanities research--including collaborative digital humanities projects--will be valued as core research.

Personnel-loss risk is also mitigated because, as represented by the candidate list for the WE1S advisory board (see under *section II.d, Staffing*), WE1S has identified a rich pool of experts who might be recruited to step in should key personnel leave. In addition, other digital humanities scholars at UCSB (a long-standing center of strength in the area) and U. Miami (which has successfully proceeded with a digital-humanities cluster hire) could be recruited if need arises.

- *Risk due to complexity of project management*.

  Because WE1S involves collaboration across three universities, work by many researchers at different levels (faculty, postdocs, graduate students, and some undergraduates), an ambitious mix of intellectual and technical activities, and an extended three-year timeline, there is a risk that weak or scattered project management will result in slowdowns in the project or lack of coordination on tasks. This risk is mitigated through the hiring of lead graduate student researchers as project managers at UCSB and U. Miami (the two campuses where research assistants and other project members work on the project both in the academic years and the summers). The risk is also mitigated because WE1S's pilot project, running since 2013, has allowed the PI and co-PI's to establish effective practices for running project meetings, managing tasks (including through the use of online team collaboration platforms such as Trello and Ryver), coordinating work between institutions, collaborating on research (e.g., producing and studying topic models together), and co-authoring proposals, papers, and other documents (e.g., WE1S's accepted proposal and paper abstracts for the Digital Humanities 2017 conference in Montreal in August 2017). In addition, the risk is mitigated by the fact that the project principals have had prior experience managing large or complex projects. PI Liu was principal investigator of the UC Office of the President funded Transliteracies project on online reading from 2005-2010, a University of California "multi-campus research group" involving 11 faculty and 35 graduate students from 14 disciplines at 7 UC campuses. Co-PI Thomas, a graduate student at the time, served as Project Coordinator for Transliteracies and was a lead RA for the NEH-funded RoSE: Research-oriented Social Environment project that later evolved from Transliteracies. Co-PI Kleinman co-directed the Lexomics Project (which developed the Lexos text analysis tool and evolved it further with a NEH Digital Humanities Start-up Grant in 2015-2017); and is also co-directing the NEH-funded Archive of Early Middle English (AEME) project. Co-PI Douglass was principal investigator of the Playpower: Learning Games for Radically Affordable Computers project. (Funded in the MacArthur Foundation Humanities, Arts, Sciences and Technology Advanced Collaboratory [HASTAC] Digital Media and Learning Competition in 2009-2010, this project led to the creation of the Playpower Foundation [later Playpower Labs]). While serving as Technical Director of the Software Studies Initiative at UC San Diego, he also collaborated with Lev Manovich, the initiative's director, on several large digital media projects.

- *Risk due to a challenge to the team’s rights to use primary source documents*.

  Like other digital humanities projects focused on large-corpus collection and analysis, WE1S could face challenges from repositories or rights-holders to its ability to access primary source documents or to analyze them en masse. This risk is mitigated because WE1S adopts workflow design principles that adhere to terms of use for its sources and fair-use "non-consumptive use" law. As detailed in *sections I.e* and *II.g*, these workflow principles include: manual searching and downloading where required, no stored copies of originals, workflows running on transformed analytical data only (e.g., "bags of words" representations of documents equivalent to lists of unique words with frequency counts), and transformed analytical data that cannot be used (even in theory) to reconstruct originals.

- *Risk due to changes in technology*.

  As in the case of other digital humanities projects, WE1S works with current technical methods, protocols, tools, and platforms--all of which, individually and in composite, evolve rapidly under the ferment of changes in research infrastructure (driven by scholarly research) and broader computational infrastructure (driven by commercial, social, governmental, and other forces).

  This risk is mitigated because WE1S's technical research environment is based as much as possible on open-standards and open-source technologies. While open technologies do not eliminate disruption, they have the advantage that open communities of programmers help with update/migration problems. In addition, WE1S has identified alternatives to some of its main open source platforms--e.g., Beaker instead of Jupyter data notebooks.

- *Risk due to research "mission creep."*

  As stated in *section II*, WE1S will be expanding its main corpus, adding sub-corpora, extending the range of its analytical methodologies, and improving its technical research environment. As always when a project extensively and rapidly scales up--and especially in digital humanities projects that work with large amounts of material using new methods--there is the risk of "mission creep."

  WE1S mitigates this risk through the "triage" step it has built into its activity timeline at the end of its second academic year (see timeline in *section II.e*). At this designated time, WE1S will assess activities related to collecting new materials and/or developing analytical and technical methods. The criteria for assessment will be a combination of value for the project's core goals (as outlined in sections *I.d, Expected Audiences and Outcomes* and *II.f, Expected Outcomes and Benefits)* and "finishability." In a manner analogous to "feature lock down" in software development, WE1S will tie off its collection and analysis aims at this time, prioritizing those with the most value that can be brought across the finish line to provide expected outcomes.

- *Risk due to empirical failure to find anticipated or hypothesized results*.

  Much of modern humanities research is hermeneutical in orientation, meaning that the act of analysis--beginning even at relatively low levels of observation (e.g., "close reading" a text)--is integrated from the start in a "hermeneutic circle" establishing a feedback circuit between parts and wholes (e.g., between low-level analyses and high-level interpretations of art, history, culture, and theory). Low- and high-level intuitions constantly adjust to each other so that there cannot be failure to reach a goal (since the goal is self-adjusting).

Digital humanities projects are different because they risk incomprehensible gaps between their low-level acts of analysis (e.g., text analysis) and high-level interpretation. In the case of WE1S's method of topic modeling, for instance, there is no assured path from what the computational algorithm defines as a "topic" (a statistically correlated collection of words in a large document set) and what could persuasively be understood by a human being as a "theme." WE1S's work during its pilot-project phase with topic modeling generated from its existing corpus indicates that understandable themes are indeed apparent. Moreover, such themes appear in proportional relations to other themes (and to originating sources, localities, and times) in ways that allow for additional interpretations. However, as WE1S expands its corpus, it does face the risk that some of its interpretive goals cannot be proven empirically. For example, one of WE1S's goals, as stated in *section I.b*., is to show "that there may be other [non-mainstream] important themes, narratives, examples, metaphors, and evidence types whose role in public discourse on the humanities is unrecognized or underweighted." Another, as set forth in its expanded research and diversity goals for the Mellon grant *(section II.b.1)*, is to understand the relation between underrepresented social groups and the humanities based on both mainstream media and media addressed to, or specifically discussing, such groups. Whether such themes and relations exist or can be found based on WE1S's materials is an open empirical question.

The mitigation for this kind of risk is that WE1S's material for research is so ample, and its avenues of inquiry so multiple, that any one dead end is likely to lead to multiple avenues of alternative inquiry. In this regard, it is important to note (as argued by Stephen Ramsay in his *Reading Machines: Toward an Algorithmic Criticism*) that the goals of the digital humanities and, more broadly, cultural analytics, are not exactly the same as those of science. In the sciences, verification or disproof of hypotheses is preliminary to the generation of new hypotheses. But in the digital humanities, analytical operations are often designed from the beginning to serve not as proof-tests but as generators of branching hypotheses. WE1S may best be understood as a project that tests branching hypotheses about the public understanding of the humanities. Failure of verification down any one branch is likely to sprout, rather than nip in the bud, ways of understanding the public view of the humanities.

j. Reporting

As instructed by the Mellon Grant Reporting Guidelines, WE1S will provide interim and final reports according to the schedule specified in the Mellon Foundation's award letter. The PI will be responsible for producing the reports, assisted for financial accounting by staff at UCSB.

III. Works & Tools Cited

4Humanities.org. Home page. Accessed April 18, 2017. http://4humanities.org.

__________, "The Heart of the Matter Topic-Modeled (A Preliminary Experiment)." November 2, 2013. Accessed April 20, 2017. http://4humanities.org/2013/11/the-heart-of-the-matter-topic-modeled-a-preliminary-experiment/.

__________. "What U.S. Politicians Say About the Humanities--A Data Set and Analysis." March 1, 2016. Accessed April 20, 2017. http://4humanities.org/4humanities-research-projects/what-u-s-politicians-say-about-the-humanities/.

4Humanities.org. WhatEvery1Says (WE1S) Project. (See under WE1S.)

Academy Data Forum. Home page, 2017. American Academy of Arts and Sciences. Accessed April 19, 2017. https://www.amacad.org/content/research/dataForumList.aspx.

Algee-Hewitt, Mark, Ryan Heuser, and Franco Moretti. "On Paragraphs: Scale, Themes, and Narrative Form." Stanford Literary Lab Pamphlet 10. October 2015. Accessed October 30, 2015. https://litlab.stanford.edu/pamphlets/.

Allen, Danielle, Chris Pupik Dean, Sheena Kang, and Maggie Schein. "Humanities Craftsmanship: A Study of 30 Years of Illinois Humanities Council Grant-making -- A Report by the HULA Team. The Humanities and Liberal Arts Assessment Research Project, February 9, 2015. Accessed April 19, 2017. http://www.pz.harvard.edu/resources/humanities-craftsmanship-a-study-of-30-years-of-illinois-humanities-council-grant-making.

Allison, Sarah. "Other People's Data: Humanities Edition." *Cultural Analytics*, December 8, 2016. Accessed April 20, 2017. http://culturalanalytics.org/2016/12/other-peoples-data-humanities-edition/.

Always Already Computational: Collections as Data. Home page, 2017. Accessed April 21, 2017. https://collectionsasdata.github.io/.

American Academy of Arts and Sciences Commission on the Humanities and Social Sciences. *The Heart of the Matter: The Humanities and Social Sciences for a Vibrant, Competitive, and Secure Nation*. Cambridge, MA: American Academy of Arts and Sciences, 2013. Accessed April 19, 2017. http://www.humanitiescommission.org/_pdf/hss_report.pdf.

Apache Taverna (Taverna Workflow System). Home page. Apache Software Foundation, 2014-2016. Accessed February 2, 2017. https://taverna.incubator.apache.org/.

Archive of Early Middle English (AEME). Development site home page, 2017. Accessed June 21, 2017. http://scottkleinman.net/aeme-dev/.

Association of American Universities (AAU). Home page. Accessed April 21, 2017. https://www.aau.edu/.

Authors Guild et al. v. Google Inc., 804 F.3d 202 (2d Cir. 2015).

Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).

Bate, Jonathan. *The Public Value of the Humanities*. London: Bloomsbury, 2011.

Belfiore, Eleonora, and Anna Upchurch, ed. *Humanities in the Twenty-First Century: Beyond Utility and Markets*. Basingstoke: Palgrave McMillan, 2013.

Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55.4 (April 2012): 77-84. doi: 10.1145/2133806.2133826.

Bod, Rens. *A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present*. Oxford University Press, 2013.

de Bolla, Peter. *The Architecture of Concepts: The Historical Formation of Human Rights*. New York: Fordham University Press, 2013.

California State University, Northridge (CSUN). "Diversity Initiatives at California State University Northridge." N. d. Accessed April 21, 2017. http://www.csun.edu/sites/default/files/CSUN%20Diversity%20Initiatives.pdf.

__________. HSI Pathways to the Professoriate. Home page, n. d. Accessed April 21, 2017. http://www.csun.edu/humanities/pathways-professoriate.

Chronicling America: Historic American Newspapers. Home page, n. d. Library of Congress. Accessed April 20, 2017. http://chroniclingamerica.loc.gov/.

Clawson, James. "Who’s Afraid of Topic Modeling? Proposing a Collaborative Workflow (with Virginia Woolf)." jmclawson.com. Accessed April 20, 2017. http://jmclawson.com/topickit/Chicago-pdf.pdf.

Clement, Tanya E. "Towards a Rationale of Audio-Text." *Digital Humanities Quarterly* 10.2 (2016). Accessed April 19, 2017. http://www.digitalhumanities.org/dhq/vol/10/3/000254/000254.html.

Congress.gov. Home page, n. d. Library of Congress. Accessed April 20, 2017. https://www.congress.gov/.

Cordell, Ryan. "Reprinting, Circulation, and the Network Author in Antebellum Newspapers." American Literary History 27.3 (2015): 417-445. Accessed February 8, 2017. https://muse.jhu.edu/article/590710/pdf.

Cultural Analytics (journal). “About CA.” Accessed March 20, 2017. http://culturalanalytics.org/about/about-ca/.

dfr-browser. (See Goldstone, Andrew.)

Docker. Home page, 2017. Docker, Inc. Accessed April 30, 2017. https://www.docker.com/.

Fedora. Home page, n. d. DuraSpace. Accessed April 30, 2017. http://fedorarepository.org/.

Gallé, Matthias, and Matías Tealdi. “Reconstructing Textual Documents from n-grams.” *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*: 329-338. New York: ACM, 2015. Accessed June 3, 2017. doi: 10.1145/2783258.2783361.

Goldstone, Andrew. dfr-browser, v. 0.8a. Home page, June 8, 2016. Accessed March 22, 2017. http://agoldst.github.io/dfr-browser/.

Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45.3 (2014): 359-84.

Harpham, Geoffrey Galt. *The Humanities and the Dream of America*. Chicago: University of Chicago Press, 2011.

HathiTrust Digital Library. "Non-Consumptive Use Research Policy." February 20, 2017. Accessed May 27, 2017. https://www.hathitrust.org/htrc_ncup.

Hispanic Association of Colleges and Universities (HACU). " HACU Member Hispanic-Serving Institutions (HSIs)." N. d. Accessed April 21, 2017. https://www.hacu.net/assnfe/CompanyDirectory.asp?STYLE=2&COMPANY_TYPE=1%2C5.

__________. " Hispanic-Serving Institution Definitions." N. d. Accessed April 21, 2017. http://www.hacu.net/hacu/HSI_Definition.asp.

Hispanic Serving Institutions (HSI). See above under Hispanic Association of Colleges and Universities (HACU), "HACU Member Hispanic-Serving Institutions (HSIs)" and "Hispanic-Serving Institution Definitions."

History of Humanities (journal). Edited by Rens Bod, Julia Kursell, Jaap Maat, Thijs Weststeijn. University of Chicago Press. Available online at: http://www.journals.uchicago.edu/toc/hoh/current.

HathiTrust Research Center. Home page, n. d. Accessed April 18, 2017. https://analytics.hathitrust.org/.

Humanities and Liberal Arts Assessment Project (HULA). Home page, 2016. Harvard Graduate School of Education. Accessed April 19, 2017. http://www.pz.harvard.edu/projects/humanities-liberal-arts-assessment-hula.

Humanities Commons. Home page, 2016. Accessed April 30, 2017. https://hcommons.org/.

Humanities Indicators. Home page, 2016. American Academy of Arts and Sciences. Accessed April 18, 2017. http://www.humanitiesindicators.org/.

Hutner, Gordon, and Feisal G. Mohamed, ed. *A New Deal for the Humanities: Liberal Arts and the Future of Public Higher Education*. New Brunswick, NJ: Rutgers University Press, 2016.

Hypothes.is. Home page, n. d. Accessed April 20, 2017. https://hypothes.is/.

Gil, Yolanda, and Daniel Garijo. "Towards Automating Data Narratives." *Proceedings of the Twenty-Second ACM International Conference on Intelligent User Interfaces, Limassol, Cyprus, February 2017*. ACM, 2017. doi: 10.1145/3025171.3025193.

GitHub. Home page. Github, Inc., 2017. Accessed January 30, 2017. https://github.com/.

International Image Interoperability Framework (IIIF). Home page, n. d. Accessed April 19, 2017. http://iiif.io/#plug-%E2%80%99n%E2%80%99-play-software.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. Champaign, IL: University of Illinois Press: 2013.

JSTOR. Home page, n. d. ITHAKA. Accessed April 20, 2017. https://www.jstor.org/.

JSTOR Text Analyzer, beta version. Home page, n. d. JSTOR / ITHAKA. Accessed April 20, 2017. https://www.jstor.org/analyze/.

Jupyter Notebooks. (See Project Jupyter.)

Kabaservice, Geoffrey. "The Birth of a New Institution: How Two Yale presidents and Their Admissions Directors Tore Up the “Old Blueprint” to Create a Modern Yale." Yale Alumni Magazine, December 1999. http://archives.yalealumnimagazine.com/issues/99_12/admissions.html.

Kepler [computer software], v. 2.5. Home page. N. d. Accessed February 2, 2017. https://kepler-project.org/.

LexisNexis Academic. Content databases accessible through institutional subscriptions via the Web sites of WE1S's campus library systems. Includes 2,926 full-text newspapers and newspaper-like publications (plus additional magazine, newsletter, newswire, and other sources). Title list available as csv file LNAUSA_Open_URL.csv, April 14, 2017. Accessed April 20, 2017. http://amdev.net/content_reports/LNAUSA_Open_URL.csv.

Lexomics Project. Home page, n. d. Wheaton College. Accessed April 20, 2017. http://wheatoncollege.edu/lexomics/.

Lexos, v. 3.0. Home page. Lexomics Research Group, Wheaton College. Accessed January 14, 2017. http://lexos.wheatoncollege.edu/.

Long, Hoyt and Richard Jean So. “Literary Pattern Recognition: Modernism between Close Reading and Machine Learning.” *Critical Inquiry* 42 (2016): 235-267.

MALLET (Machine learning for Language Toolkit). (See McCallum, Andrew Kachites.)

McCallum, Andrew Kachites. MALLET (Machine Learning for Language Toolkit). Home page, 2002. Accessed May 28, 2017. http://mallet.cs.umass.edu.

Manovich, Lev, Jeremy Douglass, and William Huber. "Understanding Scanlation: How to Read One Million Fan-Translated Manga Pages." *Image & Narrative* 12.1 (2011). Accessed April 20, 2017. http://www.imageandnarrative.be/index.php/imagenarrative/article/view/133.

Mathae, Katherine Bailey, and Catherine Langrehr Birzer, ed. *Reinvigorating the Humanities: Enhancing Research and Education on Campus and Beyond*. New York/Washington, D. C.: American Association of Universities, 2004. Accessed April 19, 2017. https://eric.ed.gov/?id=ED505820.

Meandre. Home page, n. d. Accessed April 18, 2017. http://www.seasr.org/meandre/.

Mitra, Bhaskar. "A Simple Introduction to Word Embeddings." Microsoft, Bing Services. Slideshare, April 5, 2016 (slideshow). Accessed April 13, 2017. https://www.slideshare.net/BhaskarMitra3/a-simple-introduction-to-word-embeddings.

Mohr, John, and Petko Bogdanov. "Introduction–Topic Models: What They Are and Why They Matter." *Poetics* 41.6 (2013): 545-769. Accessed April 18, 2017. http://www.sciencedirect.com/science/article/pii/S0304422X13000685.

National Humanities Alliance. Home page, 2015. Accessed April 19, 2017. http://www.nhalliance.org/.

New York Times. "Top Colleges Doing the Most for the American Dream." May 25, 2017. Accessed June 4, 2017. https://www.nytimes.com/interactive/2017/05/25/sunday-review/opinion-pell-table.html.

Nussbaum, Martha C. *Not for Profit: Why Democracy Needs the Humanities*. Princeton, NJ: Princeton University Press, 2016.

Open Science Framework. Home page, n. d. Center for Open Science. Accessed April 19, 2017. https://osf.io/.

Piper, Andrew. "Fictionality." *Cultural Analytics*. December 20, 2016. Accessed May 29, 2016. doi: 10.22148/16.011.

Playpower Foundation. Home page, n. d. Accessed June 21, 2017. http://playpower.org/.

Project Jupyter (Jupyter Notebooks). Home page, April 13, 2017. Accessed April 19, 2017. http://jupyter.org/.

ProQuest. Content databases including (in various overlapping packagings) News & Newspapers, Historical Newspapers, Ethnic NewsWatch, Black Newspapers, U.S. Hispanic Newsstand, and GenderWatch accessible through institutional subscription via the Web sites of WE1S's campus library systems. Includes full-text access to198 newspapers, 44 historical newspapers, plus magazines. See Proquest, "News & Newspapers," n. d. Accessed April 20, 2017. http://www.proquest.com/libraries/academic/news-newspapers/.

PROV. (See W3C [World Wide Web Consortium]. "PROV-Overview." April 30, 2013. Accessed February 2, 2017. https://www.w3.org/TR/prov-overview/.)

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. Urbana-Champaign, IL: University of Illinois Press, 2011.

R-Shief. Home page, 2016. Accessed September 11, 2016. http://r-shief.org/.

Rockwell, Geoffrey. "2016 Chicago Colloquium on Digital Humanities and Computer Science" (notes on the conference). *philosophi.ca* (blog). Accessed April 20, 2017. http://philosophi.ca/pmwiki.php/Main/2016ChicagoColloquiumOnDigitalHumanitiesAndComputerScience.

RoSE (Research-oriented Social Environment). Home page. University of California, Santa Barbara, 2012. Accessed July 14, 2012. http://rose.english.ucsb.edu/.

Ryver. Home page, n. d. Accessed June 20, 2017. https://ryver.com/.

Schmidt, Benjamin M. "Word Embeddings for the Digital Humanities." *Bookworm* (blog), October 25, 2015. Accessed April 13, 201i7. http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html. [*Note:* there are scripts running on this page that may cause a browser to freeze temporarily. If a browser dialogue pops up, choose to "continue" until the page fully loads.]

Schiffrin, Anya, and Ethan Zuckerman. "Can We Measure Media Impact? Surveying the Field." *Stanford Social Innovation Review*, Fall 2015. Accessed June 1, 2017. https://ssir.org/articles/entry/can_we_measure_media_impact_surveying_the_field.

SEASR (Software Environment for the Advancement of Scholarly Research). Home page, n. d. Accessed April 18, 2017. http://www.seasr.org/.

Small, Helen. *The Value of the Humanities*. Oxford: Oxford University Press, 2013.

Smith, Sidonie Ann. *Manifesto for the Humanities: Transforming Doctoral Education in Good Enough Times*. Ann Arbor: University of Michigan Press, 2016.

So, Richard Jean, and Hoyt Long. "Network Analysis and the Sociology of Modernism." *Boundary 2*, 40.2 (2013):147-182. Accessed April 19, 2017. http://boundary2.dukejournals.org/content/40/2/147.full.pdf+html.

Society for the History of the Humanities. Home page, 2017.Accessed April 19, 2017. http://www.historyofhumanities.org/.

Sunlight Foundation. Home page, n. d. Accessed April 20, 2017. https://sunlightfoundation.com/.

Text Encoding Initiative (TEI). Home page, July 19, 2016. Accessed April 19, 2017. http://www.tei-c.org/index.xml.

Transliteracies: Research in the Technological, Social, and Cultural Practices of Online Reading (University of California Multi-Campus Research Group project, 2005-2010). Principal investigator, Alan Liu. Home page, n. d. Accessed June 21, 2017. http://transliteracies.english.ucsb.edu.

Trello. "Tour" (information) page, 2017. Accessed June 20, 2017. https://trello.com/tour.

Ubois, Smiljana Antonijievic, and Ellysa Stern Cahoy. "Supporting Humanists' Digital Workflow." (See notes on talk by Rockwell.)

Underwood, Ted. "Topic Modeling Made Just Simple Enough." *The Stone and the Shell* (blog), April 7, 2012. Accessed February 15, 2017. http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

University of Miami. *Fact Book 2016-17 (Fall 2016 Fact Book)*. 2017. University of Miami. Accessed April 21, 2017. http://www.miami.edu/index.php/Fact_Book_2016-2017.

U. S. Department of Education. "Developing Hispanic-Serving Institutions Program - Title V." N. d. Accessed April 21, 2017. https://ed.gov/programs/idueshsi/index.html?src=rt.

U. S. Government Publishing Office. Home page, n. d. Accessed April 20, 2017. https://www.gpo.gov/.

The Voice of the Shuttle. Home page, n. d. Edited by Alan Liu. University of California, Santa Barbara. Accessed April 30, 2017. http://vos.ucsb.edu/.

WE1S (WhatEvery1Says) Project Developers' Web site. http://4humwhatevery1says.pbworks.com. (Current synopsis of project presented on a public-facing Web site: http://4humanities.org/category/whatevery1says/).

__________. "How Public Media in the US and UK Compare in Their Terminology for the Humanities." August 3, 2015. Accessed May 27, 2017. http://4humwhatevery1says.pbworks.com/w/page/98623971/How%20Public%20Media%

20in%20the%20US%20and%20UK%20Compare%20in%20Their%20Terminology%20For%20the%20Humanities.

__________. Making the Humanities Public (undergraduate collaborative research group project). Home page, December 18, 2016. Accessed April 21, 2017. http://liucrgs.pbworks.com.

__________. "Topic Modeling Systems and Interfaces." November 12, 2016. Accessed April 20, 2017. http://4humwhatevery1says.pbworks.com/w/page/104256241/Topic%20Modeling%20Systems%20and%20Interfaces.

Whitehouse.gov (The White House). Home page, n. d. Accessed April 20, 2017. https://www.whitehouse.gov/.

Wilkens, Matthew. "Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction." *Cultural Analytics*, November 1, 2016. Accessed March 1, 2017. Article doi: 10.22148/16.009. Dataset doi: 10.7910/DVN/EXPXYT.

Wings. Home page. May 8,2016. Accessed February 2, 2017. http://www.wings-workflows.org/.

Yang, Henry. " A Message on Diversity from Chancellor Henry T. Yang." University of California, n. d. Accessed April 10, 2017. http://diversity.evc.ucsb.edu/message.from.the.chancellor/.

Zenodo. "About Zenodo" (home page), n. d. CERN (European Organization for Nuclear Research). Accessed April 30, 2017. http://about.zenodo.org/.

Budget Narrative

The budget documents that are the basis for the narrative below include the main budget spreadsheet formatted according to the Mellon template (referred to below as "Main Budget") and three worksheets in the Appendix (referred to as "UCSB," "CSUN," and "UM") showing more detail on expenses at WE1S's grant home and two subgrantees. (The worksheets for the subgrantees are formatted as received from, and approved, by each campus.)

WE1S seeks a grant of $1,099,656 from the Mellon Foundation for a three-year timeline of research, development, and output activities from October 1, 2017, to September 30, 2020. Of this total, $680,960 supports project work at the grant home, the University of California, Santa Barbara (UCSB). Another $246,510 and $172,186, respectively, support work at two subgrantee institutions: California State University, Northridge (CSUN), and University of Miami (UM).

Project costs are generally spread evenly across the three project years, with exceptions due to specific institutional, personnel, and timing issues as described below. Another exception is that expenses indexed to salaries (salary supplements, benefits, and course buyouts) rise slightly each year on a predicted slope of between 2% and 3% depending on the institution and type of expense.

I. Salaries

a. Salary Supplements for Project Principals and Other Faculty

The project PI and co-PIs (Professors Alan Liu and Jeremy Douglass at UCSB; Professor Scott Kleinman at CSUN; and Professor Lindsay Thomas at UM) are budgeted for summer salary supplements during each project year equivalent to "one month" of each individual's annual salary, except in the case of PI Liu, who is budgeted for less than "one month" due to the disproportionate size of his salary (UCSB lines 25, 27; CSUN line 9; UM line 6).[22] The exact calculation of "summer one months" (a common unit-term for salary supplements), and of expected yearly increases, varies by institution. At UCSB, a summer one month amounts to 1/9th of salary; at CSUN it amounts to 1/8th; and at UM it is 1/9th.

Base salaries in project year 1 for the principals (on which salary supplements and benefits are figured) are as follows: Liu ($209,100, incremented by a predicted 2% each successive year), Douglass ($85,476, incremented 2% each year), Kleinman ($91,684, incremented 1.03% each year), Thomas ($70,040, incremented 3% each year).

Professor Mauro Carassai at CSUN is budgeted for a fixed summer salary supplement each year of $8,000 for helping lead the project's summer research camps at CSUN (CSUN line 13). In addition, two other CSUN faculty members yet to be determined are budgeted for fixed summer salary supplements of $2,000 each year to participate in the summer research camps (CSUN lines 15).

b. Course Releases ("course buyouts")

The project PI and co-PIs are budgeted for course releases. Such releases are implemented through "course buyouts" that provide their institutions with funding that can be used for the salary of replacement instructors. Depending on the way each institution handles course buyouts, these are reported either as a single item for salary (in the worksheets for UCSB and UM) or as both salary and benefits items (in the worksheet for CSUN).

Alan Liu and Jeremy Douglass at UCSB are budgeted for one course release per year. Lindsay Thomas receives the same total number of releases on a different schedule. She takes no release in project year 1, due to the conditions of her appointment then as a fellow at UM's humanities institute. Instead, she takes two course releases in the project's second year, plus one in the third year. Because Scott Kleinman at CSUN has twice the normal course load as the other project principals (he teaches eight courses per year), he receives two course releases annually (for a total of six) to allow him to devote equivalent time to the project.

[22] The UM worksheet as received from, and approved by, U. Miami aggregates summer 1-month salary supplements with academic-year course releases (course buyouts) for co-PI Lindsay Thomas on line 6. It aggregates benefits on line 11.

Course buyouts are detailed in UCSB lines 26, 28; CSUN line 10; and UM line 6.[23] Though determined through formulas set at each institution, buyouts for these particular personnel amount to the same 12.5% of annual salary and benefits. At UCSB, the policy is that a buyout covers 12.5% of salary for faculty teaching a normal course load of four courses per year (which applies to Alan Liu and Jeremy Douglass). At CSUN, one course buyout costs 1/8th or 12.5% of a faculty member's salary (since faculty teach eight courses per year). At UM, one course buyout also costs 12.5% of salary.

c. Staff Salaries

Thomas Padilla, Humanities Data Curator on the staff of the UCSB Library, is budgeted for 10% of his regular salary (UCSB line 29), based on his salary in project year 1 of $77,250. With the consent of the UCSB Library, he is committing 10% of his regular job effort as an embedded member of WE1S. (This 10% figure is not a supplement on top of his regular salary. Non-faculty staff at UCSB are not allowed to receive salary supplements as project members.)

d. Postdoctoral Faculty Fellow Salaries

Two "postdoctoral faculty fellow" positions to be filled through open recruitment are budgeted for UCSB in project years 2 and 3. (Year 1 lies fallow because of the lead time required to advertise for and fill the positions.) These postdoctoral faculty fellows each receive annual compensation in year 2 of $39,235 and year 3 of $40,020 in salaries sourced from the project grant (UCSB lines 30, 31, 40, 41). Each also receives additional annual salary of $19,939 in year 2 and $20,337 in year 3 in the form of cost share committed by UCSB for serving as lecturers for three courses per year (cost share not included in the budget for the Mellon). The total annual compensation per position is thus approximately $60,000 each year, structured as follows:

- appointment as a postdoc on the UCSB postdoc salary scale at 100% time for 3 months and 67% time for 9 months (funded from the project grant);
- appointment as a lecturer on the UCSB lecturer salary scale at 33% time for 9 months (funded by UCSB).

e. Graduate Student Researcher (GSR) Salaries at U. Miami

There is one position in the budget each project year at UM for a "graduate student researcher" (GSR) funded during the academic year (see UM line 7). However, in project year 1 this position is budgeted just for Spring semester. Because WE1S will not learn if its Mellon proposal is approved until September 2017 (after the start of Fall semester at UM), a GSR can only be recruited for Spring semester.

f. Research Assistant (RA) Salaries

The project budget includes positions for a variety of graduate, and some undergraduate, research assistants (RAs) to be hired on hourly wages.

[23] (See note above.)

UCSB

UCSB line 32: Summer research camps employ 12 graduate-student RAs each summer at $15.50/hour (the normal RA rate for humanities departments). Each RA works up to 120 total hours (approximately six weeks at an average of 20 hours of work per week).

UCSB line 33: During the academic years, three graduate-student or undergraduate RAs (depending on recruitment of available students to the positions) each work up to a total of 300 hours, providing assistance to, and participating in, project activities.

UCSB lines 34: Recruited as needed during summers and academic years, three students with computer programming skills serve as research assistants at $24/hour for a total of 97 hours each during each project year. Supplementing the programming and other technical skills of the project's principals, they help with the project's technological development. (The hourly wage for such students is higher than for other hourly RAs because of competition for their skills on campus. Past digital humanities projects at UCSB have found it necessary to offer higher hourly pay for such students.) These students are either graduate students or undergraduates depending on availability. (Past digital humanities projects at UCSB have benefited, for example, from participation by graduate students in the campus's Media Arts & Technology program and undergraduates in Computer Science.)

CSUN

CSUN line 17: Summer research camps employ 10 RAs (in a mix of approximately 8 graduate students and 2 undergraduates) each summer at $15.50/hour. RAs each work up to 120 total hours per summer (approximately six weeks at an average of 20 hours of work per week). While most RAs will be graduate students, CSUN will also seek the participation of undergraduates as part of the project's diversity aims.

UM

UM line 8: During academic years, three graduate students serve as RAs at $15/hour for 300 hours each, assisting in such project work as data collection, cleaning, and analysis.

Salaries Total: $534,227

Subtotals by campus:

- UCSB $534,227
- CSUN $192,070)
- UM $146,413

II. Benefits

Benefits included in the project budget are indexed to salaries according to the formula of each institution. The underlying benefit rates are as follows.

UCSB

- *UCSB lines 37, 38*: Benefits for summer salary supplements for PI and co-PI:

  Alan Liu (PI), Professor @ 6.60%

  Jeremy Douglass (co-PI), Asst. Professor @ 12.80%

- *UCSB line 39*: Benefits for Thomas Padilla (project Sustainability and Usability Advisor) @ 48.07%
- *UCSB lines 40, 41*: Benefits for postdoctoral faculty fellows @ 18.30%
- *UCSB lines 42, 43, 44*: Benefits for research assistants @ 3.10%

CSUN

- *CSUN lines 22, 23, 24, 25, 26*: Benefits for co-PI, other faculty, research assistants, and as part of course buyout costs @ 7.3%.

UM

- *UM line 11*: Aggregated benefits for co-PI (@ 26.2%) and RAs (see p. 3 of the UM worksheet: "A contribution towards health insurance is included for the full-time graduate student during the academic year based on the current rate of $2,070")

Benefits Total: $122,221

Subtotals by campus:

- UCSB $51,044
- CSUN $49,940
- UM $21,273

III. Travel

Travel for Advisory Board Conference Meeting

Travel by members of the project's advisory board to the conference meeting at UCSB planned for project year 1 is budgeted for 10 extramural participants for a total of $12,470 (UCSB, lines 48-50). This figure includes transportation (airfare, driving mileage, taxis), lodging, and meals.

1. Airfare and driving mileage costs ($4,870) are estimated on the *Travel Expenses Worksheet* included in the *Appendix*. Kayak.com was used to estimate the round-trip airfare of each individual listed among candidates for WE1S's advisory board from their home city (named in *section II.d, Staffing*). Google Maps was used to estimate the mileage for those candidates for the advisory board who are likely to drive. Probable taxi costs for those flying were added. Then the sum for all the candidate advisory board

members was averaged and multiplied by 10 to produce the estimated transportation total.

2. Lodging costs for the advisory board in Santa Barbara totaling $4,920 (UCSB line 49) are estimated based on three nights for 10 people at $164 per night. This per-night figure is the maximum allowed at UCSB based on the government per diem rate for the campus. (A check of a suitable hotel near UCSB showed a room rate of $179/night that is close to the UCSB allowed maximum.)
3. Meals for the event are budgeted for 10 people over 4 days at an estimate of $67 per day per individual for a total of $2,680 (UCSB line 50). There is no method this far in the future to provide a concrete cost estimate for this figure. It is based on the experience of the WE1S PI with Santa Barbara restaurants and is a good guess, since breakfasts will be included in the lodging, lunches will be catered fairly cheaply from UCSB campus facilities, and the only expensive meals will be dinners at restaurants, for which pre-arranged set-menu meals can be arranged for a lower than usual per-individual cost.

Travel by PI and Co-PIs

Each year, the project's principals plan to travel between the grant home and subgrantee institutions to conduct face-to-face meetings with project participants at each other's campus (including with research assistants at each institution). The airfare (between UCSB/CSUN and UM) and driving mileage (between UCSB and CSUN) are also estimated on the Travel Expenses Worksheet in the *Appendix*, where the sources of estimates are Kayak.com and Google Maps, respectively.

- The PI and co-PI based at UCSB are budgeted each year for one trip to UM (airfare, lodging, and meals) (UCSB lines 53, 54, 55); and also driving mileage for one day-trip to CSUN (UCSB line 56).
- The co-PI at CSUN is budgeted each year for one trip to UM (airfare, lodging, and meals) (CSUN lines 32, 33, 34); and also driving mileage for one day-trip to UCSB (CSUN line 37).
- The co-PI at UM is budgeted each year for one trip to UCSB, with airfare, lodging, meals, and local transportation (UM line 15).

Travel Total: $26,018

Subtotals by campus:

- UCSB $18,518
- CSUN $3,840
- UM $3,660

IV. Fellowship Stipends

UCSB is budgeted for one "graduate student fellow - stipend" position (Main Budget, line 29; UCSB detailed budget, line 63). This is a lead research assistant to be recruited from among UCSB English Department graduate students for three quarters of project year 1 (winter, spring, summer quarters) and the full years of project years 2 and 3 (fall, winter, spring, summer). Normally, such a position at UCSB would be a "graduate student researcher" (GSR) whose total compensation includes salary, tuition, and fees (benefits) for the fall through spring academic-year portion of a full year. (The inclusion of tuition and fees is necessary to fill such positions, since they would otherwise not be financially competitive with the TA or GSR positions that normally support Ph.D. students in the UCSB English Department and similar departments.) Substituting for a GSR, the "graduate student fellow" position is a workaround designed to allow UCSB cost-share to cover the tuition and fees part of the package, which according to Mellon guidelines cannot be sourced from the grant. Mellon funding is thus the source only for the equivalent of a GSR's salary, which the university will convert into an award to the student "fellow" in the form of a fellowship stipend. By agreement with UCSB's Graduate Division and English Department, such an award of an externally-sourced fellowship to a graduate student triggers a provision by which Graduate Division and the student's department together contribute cost share to cover tuition and fees.

The graduate student fellow position at UCSB is thus budgeted as follows: a stipend from UCSB sourced from the Mellon grant, plus cost share from UCSB for tuition and fees (totaling $53,853 for the three project years; not shown in budget documents for the Mellon).

However, only three quarters (winter and spring quarters, plus the following summer quarter) of such a position is budgeted for project year 1. This is owing to the timing of the Mellon grant. If the Mellon Foundation approves the project grant proposal in the cycle that leads to a decision in September 2017, then the project will start on October 1, 2017. However, it is not possible for UCSB to assign a graduate student to the "graduate student fellow" position at the start of academic year 2017-2018 based on contingent grant approval.

Fellowship Stipends Total: $58,672

Subtotals by campus:

- UCSB $58,672

V. Honoraria

The budget includes an honorarium of $1,000 each for 10 advisory board members for participation in the board's conference meeting at UCSB in project year 1 (Main Budget, line 30). The honorarium amount of $1,000 was determined by taking the median of recent honoraria offered by six centers or initiatives in the humanities at UCSB (including the campus's Interdisciplinary Humanities Center), whose directors or staff provided a sampling of their recent speaker honoraria. (See *Appendix, "Honoraria Estimates Worksheet."*)

Honoraria Total: $10,000

Subtotals by campus:

- UCSB $10,000

VI. Computer Hardware
VII. Software and Cloud Services

Expected technology costs for hardware, software, and cloud services are listed on Main Budget, lines 31-32, and detailed in the *Appendix* in a *Technology Estimates Worksheet*. As mentioned in *section II.b.2*, WE1S anticipates needing relatively minor hardware and software/cloud service upgrades for its development work. The workstations and software at the UCSB Transcriptions Center (located in the English Department), supplemented by a NAS ("networked attached storage") device and software controlled by co-PI Jeremy Douglass, provide the project with its base technological apparatus for development. This apparatus is supplemented by the workstations and laptops of individual principals and research assistants. As the project moves forward, it will need some limited additional hardware and software/cloud services to overcome processing and storage bottlenecks in development.

As itemized in the *Technology Estimates Worksheet* in the *Appendix*, WE1S budgets in the hardware category for a server (in the form of a NAS machine suitable for straightforward implementation of the WE1S Virtual Workspace System) and RAM memory upgrades. These purchases are planned for early in the project (primarily year 1).

WE1S also budgets for the combination of software and cloud-platform services listed in the *Technology Estimates Worksheet*. Software expenses include licenses for the Outwit Hub "scraping" tool for extracting downloaded documents and metadata in the plain-text format needed by WE1S. Cloud-platform expenses include per time/bandwidth/memory rates for use of such platforms as Amazon AWS designed for big-data scale processing. Expenses for software and cloud platform services are tapered more gradually through the three project years by contrast with hardware expenses, which are front-loaded at the beginning of the project.

These hardware and software costs include a 10% contingency figure to allow WE1S to adjust to possible future cost rises or changes in available technology.

Hardware Total: $2,415

Software/Cloud Services Total: $1,765

Subtotals by campus:

- UCSB $2,415 (hardware), $1,765 (software/cloud services)

VII. Supplies for Board of Advisors Meeting
IX. Supplies for meetings at U. Miami and CSUN

Supplies for meetings and other activities (UCSB lines 67, 68; CSUN line 42; UM line 16) will be purchased as needed. They are expected to include: upgraded omni-directional microphones for online project meetings, whiteboards and other presentation or collaboration materials, and photocopied handouts. Meetings include the advisory board conference meeting at UCSB during project year 1; project team meetings conducted on each campus or through remote conferencing between campuses; meetings during the summer research camps, and meetings when the PI and co-PIs travel to each other's campuses.

Supplies for Board of Advisors Meeting Total: $1,400

Subtotals by campus:

UCSB $1,400

Supplies for meetings at U. Miami and CSUN Total: $2,420

Subtotals by campus:

- UCSB $920
- CSUN $660
- UM $840

X. Recruitment Costs for Postdoctoral Faculty Fellows

WE1S includes in its budgets (Main Budget, line 35) a total of $2,000 for recruiting postdoctoral faculty fellows. These expenses go toward placing ads in the job information lists of professional scholarly organizations such as the Modern Language Association (which currently charges about $595 for a job ad). WE1S will place ads in three or more such venues as needed. (Recruitment for a postdoc in the third year will only be needed if a postdoc does not stay two years, due, for example, to getting a tenure-track job elsewhere.)

Recruitment Costs for Postdocs Total: $2,000

Subtotals by campus:

- UCSB $2,000

XI. Unspent Grant Funds Policy

According to the UCSB Sponsored Projects Office, standard campus practice is that if there are residual grant funds at the end of the project period, the unspent funds are returned to the sponsor.

Budget Spreadsheet

Appendices

1. Detailed Budget Worksheets for UCSB, CSUN, and UM
2. Glossary for WE1S Technical Environment
3. Draft Job Descriptions
4. Travel Expenses Worksheet
5. Technology Estimates Worksheet
6. Honoraria Estimates Worksheet
7. Curriculum Vitae