The WhatEvery1Says (WE1S) Project
— A Prospectus

The following prospectus for the WhatEvery1Says (WE1S) Project is a synopsized, revised version of the original proposal narrative that the WE1S team included in the grant proposal it submitted in 2017 to the Scholarly Communications Program of The Andrew W. Mellon Foundation. This shorter version of the proposal narrative is designed to bring forward the project's main ideas, goals, and plans.

—— *Alan Liu, Jeremy Douglass, Scott Kleinman, and Lindsay Thomas (October 1, 2017)*

Table of Contents

I. Fact Sheet

Participating Institutions

- University of California, Santa Barbara (UCSB)
- California State University, Northridge (CSUN) (subgrantee)
- University of Miami (UM) (subgrantee)

Team

- *Principal Investigator:* Alan Liu (Professor of English, UCSB)
- *Co-PIs:* Jeremy Douglass (Assistant Professor of English, UCSB), Scott Kleinman (Professor of English, CSUN), Lindsay Thomas (Assistant Professor of English, UM)
- *Other participating faculty:* Mauro Carassai (CSUN)
- *Sustainability and Usability Adviser:* Greg Janée (Director, Data Curation Program, UCSB Library; and Developer, UC Curation Center at California Digital Library)
- *[To be recruited:]* Two Postdoctoral Scholars/Lecturers (postdoctoral researchers who also teach some courses each year) for project years 2018-2019 and 2019-2020
- *[To be recruited:]* Graduate student project managers
- *[To be recruited:]* Graduate student and undergraduate research assistants, including those participating in WE1S's "summer research camps"

Funding Support

- The Andrew W. Mellon Foundation, $1.1 million
- Additional support from UCSB in the form of additional salaries and benefits for the hybrid postdoc/lecturer positions and for the graduate-student project manager at UCSB.

Timeline

- Three years, October 1, 2017 to September 30, 2020.

Main Humanities Research Goals

- Use digital-humanities computational methods to study public discourse on the humanities in journalistic media and other sources at large data scales. Model the main themes, frames of discussion, and narratives (even "memes") of such discourse; the relations between them; and also unexpected themes or relations. For example, how do journalists, politicians, business people, scientists, parents, students, university administrators, professors, writers, artists, and others typically talk about the humanities? How do the national and local, or policy and personal facets of "what everyone says" about the humanities intersect with each other (for example, when a

legislator discusses education policy, or a parent or teacher discusses the future with a student)?

- Study how public discourse on the humanities compares across states, nations, and regions of the world. To begin with, how much talk about the humanities is there in the media in California as opposed to in New York, in the U.S. by contrast with the U.K., and in the North and West of the world by comparison with the southern hemisphere or mid and far east?

- Study the way racial, ethnic, gender, and other groups are positioned by the media, or position themselves in the media, in relation to the humanities versus more "practical" career choices, especially during first-generation immigrant or first-generation-to-college phases of a group's social trajectory. Just one line of inquiry going beyond the stereotypes: if a first-generation immigrant student is told by parents and society to major in science, engineering, pre-medicine, pre-law, or pre-business; yet the cultural and personal identity of that student is vested in a deep humanities and arts heritage; then where does that excess "humanity" go and how is it expressed and cultivated?

Main Digital Humanities Goals

- Design and assemble a representative corpus (in this case, of public discourse on the humanities) with attention to principles of corpus "representativeness."

- Advance methods for open, reproducible data workflow in the digital humanities, including methods for
  - tracking the provenance and workflow of sources, tools, and processes;
  - implementing workflows of data collection, preparation, analysis, and output workflows in chained series of executable (but customizable) data "notebooks" (Jupyter notebooks).

- Advance the sustainability of digital humanities workflows and outputs by using "containerized" (virtualized) data analysis platforms that can be distributed and be deposited in institutional or other repositories.

Main Public Goals (addressed to the public or to humanities advocates addressing the public)

- Provide a richer stock of themes, narratives, examples, and evidence types that can be drawn upon in discussing the humanities, 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 life or career is about).

- Provide recommendations for humanities advocacy in the form 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).

- Provide resource "kits" of themes, examples, and evidence for journalists, scholars, and administrators to draw on in discussing the humanities; or for students to draw on as they consider choosing a major and discussing it with parents.

II. 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 digital textual form beginning circa 1981. Previously engaged in a small-scale pilot project for this purpose, WE1S was in 2017 awarded Mellon Foundation funding of $1.1 million 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 Principal Investigator Alan Liu (UCSB) 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—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*. 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 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. The main examples of scientific data workflow systems, which are even more powerful and complex, are Apache Taverna, Kepler, and Wings. By contrast with such systems, WE1S's combined Workflow Management System and Virtual Workspace System 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 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, cleaning, analysis, and visualization optimized for the individual researcher or small-team digital humanities project.[2] 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

---

[2] 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.

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).[3] 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[4]), 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."[5] 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).[6] WE1S's development of methods for declaring, annotating, and 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.

---

[3] 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.

[4] 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.

[5] 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".

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

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.[7] 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.

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) that include the phrases "humanities," "liberal arts," and (in the United Kingdom and Commonwealth nations, "the arts").[8] For example, WE1S's pilot project 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.[9] 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

---

[7] 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.

[8] 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)*.

[9] 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).

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.[10] 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 will require storing other non-consumptive, non-reconstructable document data in the same fashion.

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.

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

---

[10] "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").

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 might send human interpreters back to access and read some of the original articles

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 scor
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 lesse
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.

contributing to that topic.[11]

---

[11] 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. 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, ranked articles associated with a topic, and the changing importance of
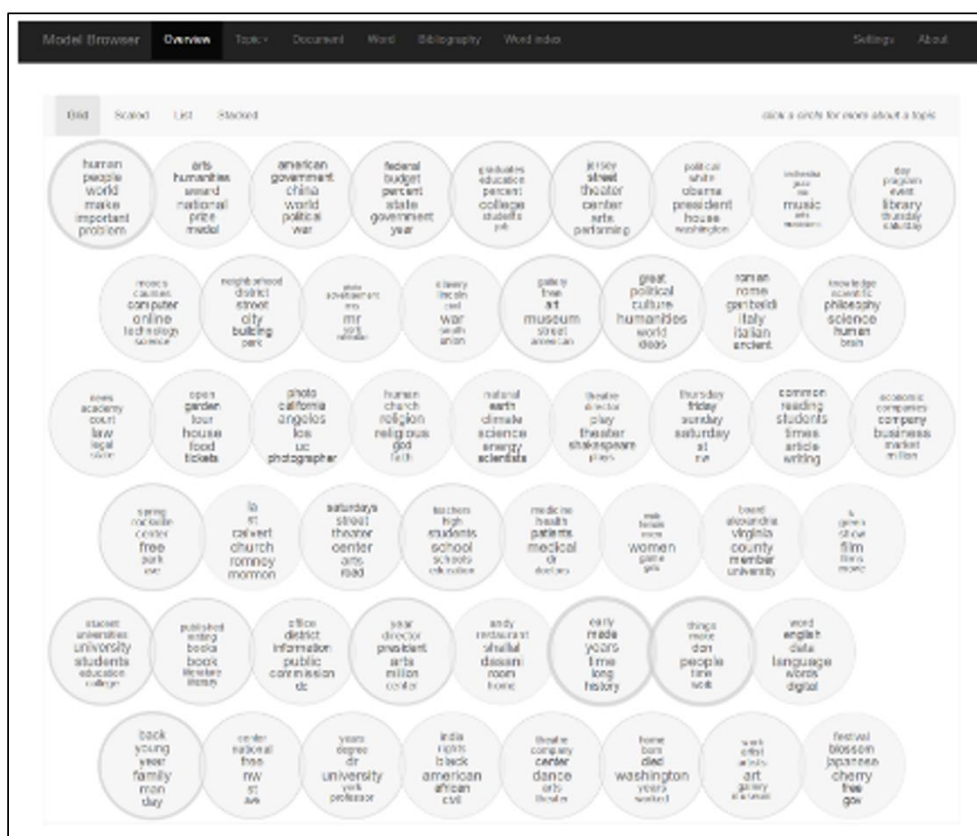

Figure 2: Dfr-browser interface showing one of several dynamic, interactive views of a topic model generated by WE1S during its pilot project.

a topic in the total document set over time.

---

[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.

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.

f. Technical Methods

WE1S has developed 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. 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:

1. Corpus Assembly and Preparation System

WE1S collects as "plain text" (the most tractable format for computational analysis) the materials from its journalistic sources via databases 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, 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.

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/whatevery1says/manifest/blob/master/we1s-manifest-schema-1.1.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.

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

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

  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).

- 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

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

  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.

- 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

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

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).

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.

III. 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. In this phase, the project focused on developing research questions and goals, research methods, and technical methods. 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. The developers' Web site for the pilot project, which hosts extensive planning documents, progress reports, and meeting notes, is at http://4humwhatevery1says.pbworks.com. A public facing Web site for the project has recently been started at http://we1s.ucsb.edu.

Preliminary analyses of experimental topic models produced during the pilot project have shown the promise of the topic modeling approach. WE1S discovered, for example, 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 above—*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).* 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

With funding from the Mellon Foundation, 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).

Specifically, WE1S's plans 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, 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, 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)[14];

- Reports and publications by scholarly and professional associations as well as grant agencies and foundations[15];

- 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);

---

[14] 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."

[15] 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)."

- 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.)

2. Improvement of Research and Technical Methods

Currently, WE1S's research methods and their technological implementation 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; 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). 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.

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

- *Improve methods for quick, iterable corpus assembly.* The goal is to allow WE1S researchers rapidly and flexibly to add more materials to the corpus as dictated by evolving research questions.

- *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.* WE1S will improve the integration between its Workflow Management System and Virtual Workspace System to 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.

- *Develop a public-facing front end for displaying and exploring the results of the project.* The goal is to create a Web site that combines standard site or blog functions with the ability to 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.

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 above, WE1S has identified resources (databases such as Ethnic Newswatch, Proquest Black Newspapers, and Proquest U.S. Hispanic Newsstand) in which materials can be researched 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?[16]

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

---

[16] 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.

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. WE1S's planned Summer Research Camps for graduate and undergraduate students will enhance this dimension of the project further. At UCSB, these summer research camps 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. 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, 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 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, enriched 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. 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.

IV. 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.

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/.

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/.

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

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.

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

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/.

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

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.

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/.

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/.

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.

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/.

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/.

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 subscriptionSee 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/.)

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.

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.

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.

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/.

WE1S (WhatEvery1Says). Home page, 2017. http://we1s.ucsb.edu.

_____. 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%20F
or%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%20Sy
stems%20and%20Interfaces.

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

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/.