Open, Shareable, Reproducible Workflows for the Digital Humanities: The Case of the 4Humanities.org 'WhatEvery1Says' Project

Ashley Champagne
ashleychampagne@umail.ucsb.edu
Department of English
University of California, Santa Barbara

Jeremy Douglass
jeremydouglass@english.ucsb.edu
Department of English
University of California, Santa Barbara

Scott Kleinman
scott.kleinman@csun.edu
Department of English
California State University, Northridge

Alan Liu
ayliu@english.ucsb.edu
Department of English
University of California, Santa Barbara

Jamal Russell
jamalsrussell@umail.ucsb.edu
Department of English
University of California, Santa Barbara

Lindsay Thomas
lindsaythomas@miami.edu
Department of English
University of Miami

Introduction

This panel reports on the open, shareable, and reproducible workflow methodology for digital humanities research developed by the 4Humanities.org "WhatEvery1Says" (WE1S) project. WE1S is topic modeling a large corpus of articles related to the humanities in newspapers, magazines, and other media sources in the U.S., U.K., and Canada from 1981 on. While the panel presents WE1S's conceptual goals and prototype experiments in using outcomes in humanities advocacy, its focus is on the technical and interpretive workflow developed by the project for humanities-oriented data work. WE1S's *manifest system* for data provenance and workflow management, its *virtual workspace manager* for integrated, containerized data manipulation and processing, and its *interpretation protocol* for how humans read topic models suggest a generalizable open approach based not on particular technologies and methods but on annotated methods. Moreover, there is a philosophical fit between such an approach and the public-facing goals of the WE1S project. WE1S is about opening public culture to view through analytics, while its DH methodology is about opening up scholarly expertise itself through shareable, transparent processes not locked into technically complex, pre-established, or large-scale research frameworks.

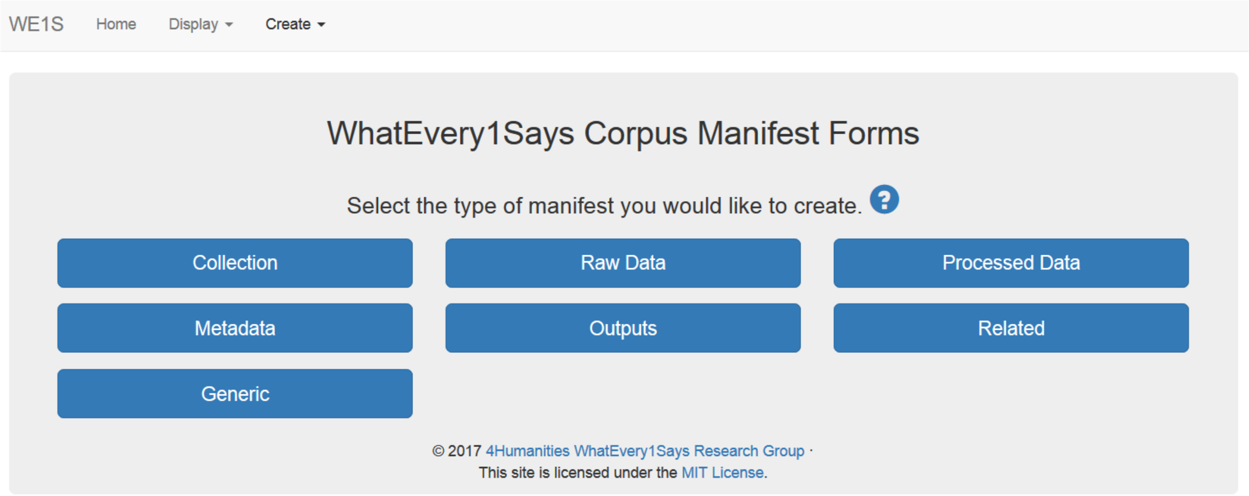
Project Research and Advocacy Goals: WE1S uses topic modeling to explore the idea of "the humanities" in public discourse. A complex concept of the kind that Peter de Bolla treated in his 2013 *The Architecture of Concepts* (his main example: "human rights"), "the humanities" as they are perceived are both tightly bunched in academic disciplines and broadly dispersed in extra-academic domains. Discussion focused on "the humanities crisis," "the decline in humanities majors," etc. creates flash points in the discourse. Yet the overall heat map of articles about the humanities, WE1S discovers, also extends into vast stretches of warm or cool discussion about humanities subjects intricately interwoven into the background of other domains of social life. Even articles so seemingly unremarkable (yet fully remarkable when we think about it) as an obituary or wedding announcement can mention the humanities as part of their *donnée*. WE1S seeks to open to view this whole conceptual architecture of "the humanities" as it exists in robust, living relation with culture at large.

Project Methodological Goals: Due to the lack of widely shared technical conventions and appropriate scholarly and publishing practices, today it is very difficult for a DH scholar to answer with documentation such questions as: *Where did you get those thousands of works in your corpus? Where did the metadata come from? What steps did you take to prepare and process the material? How many variations did you try? Where in the process was it critical for there to be "humans in the loop"?* The WE1S project addresses a growing need for ways to share and reproduce data-workflow in digital humanities research in order to make DH comparable to "open science" (see Bare, 2014). Indeed, data-intensive work in the sciences offers an especially good paradigm because of the degree to which it makes workflow and provenance management itself a thoughtful research field (i.e., research *about*, and not just tools for, managing workflow and provenance) (e.g., see Gil et al., 2007; and Garijo et al., 2012). The WE1S project is developing a technical framework that explores how the digital humanities can evolve similar, but also necessarily different, *humanities*-adapted standards of openness, shareability, and reproducibility. What amount of data and metadata, in what detail, at which processing stages, with what accompanying scripts, and so on, should be shared to support rich and persuasive scholarly discourse based on digital humanities research in the future? How will the criteria of "reproducibility, replication, and generalizability" (on the different shades of meaning of these terms in the sciences, see Bollen et al., 2015) join more traditional ideas of excellence in the humanities (e.g., "critical rigor") in the various contexts of collecting, curation, exhibition, editing, analysis, interpretation, and other work?

(1) Overview (Alan Liu). WE1S explores public thought about the humanities, especially as mediated in journalistic articles that stage a dialogic relation between leaders in government, business, universities, the arts, and others with citizens. The project's end goal is to use such research to guide humanities advocacy. But rather than create a one-off project, the WE1S group has developed a robust technical and interpretive methodology comparable (though customized for DH) to scientific data workflow management systems (e.g., Apache Taverna, Kepler, Wings), provenance tracking systems (e.g., ProvONE), and similar schemes (see Gil et al., 2007; and Bose and Frew, 2005).

(2) Manifest System for Data Provenance and Workflow Management (Scott Kleinman). The WE1S manifest schema uses a JSON Schema-based model to produce "manifests" that document the provenance of articles studied by WE1S as well as later transformations of the data, tying together scripts, stop word lists, outputs, visualizations, etc. used in project work. Manifests make the workflow

transparent and facilitate on-the-fly reiterations or adjustments--e.g., staging a subset of the WE1S corpus for topic modeling or defining variant numbers of topics. Manifests are human-readable JSON files that are highly interoperable with other systems; they can be used programmatically to drive scripts or to crosswalk information to other workflow tools or metadata frameworks. The WE1S workflow management system uses the manifest schema to generate web forms, enabling non-technical users to create and query manifests, which are stored using the same JSON-like format in its MongoDB database. The system is easy to deploy and can be adapted for other humanities research projects simply by modifying the manifest schema.



WhatEvery1Says Corpus Manifest Forms

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

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

Fig. 1: Web interface for WE1S manifest system.

(3) Virtual Workspace Manager for Integrated, Containerized Data Manipulation and Processing. (Jeremy Douglass). To address a range of computing demands from geographically distributed participants, the WE1S Workspace Manager facilitates open, reproducible DH research through a defined computing platform, a shareable online environment, integrated customizable workflows, and on-demand publishing of results. Tools for topic modeling workflows are configured on a virtual machine (a Docker container). Open data science notebooks (iPython / Jupyter) are the interface. Project templates are collections of notebooks (Python, R) chained into flows. Each new data exploration flow customizes a template, imports manifest data (from the WE1S manifest system), builds a topic model (Mallet), generates a visual browser (Andrew Goldstone's dfr-browser), publishes the browser to an interactive website, and packages a project for download and offline viewing. WE1S hosts a shared workspace online; it also runs on a laptop. Design and implementation of this virtualized, integrated workflow environment may be relevant to other DH projects, and is consonant with the philosophy of such other online or containerized integrated systems as Lexos or DH Box designed to make advanced DH research environments accessible.

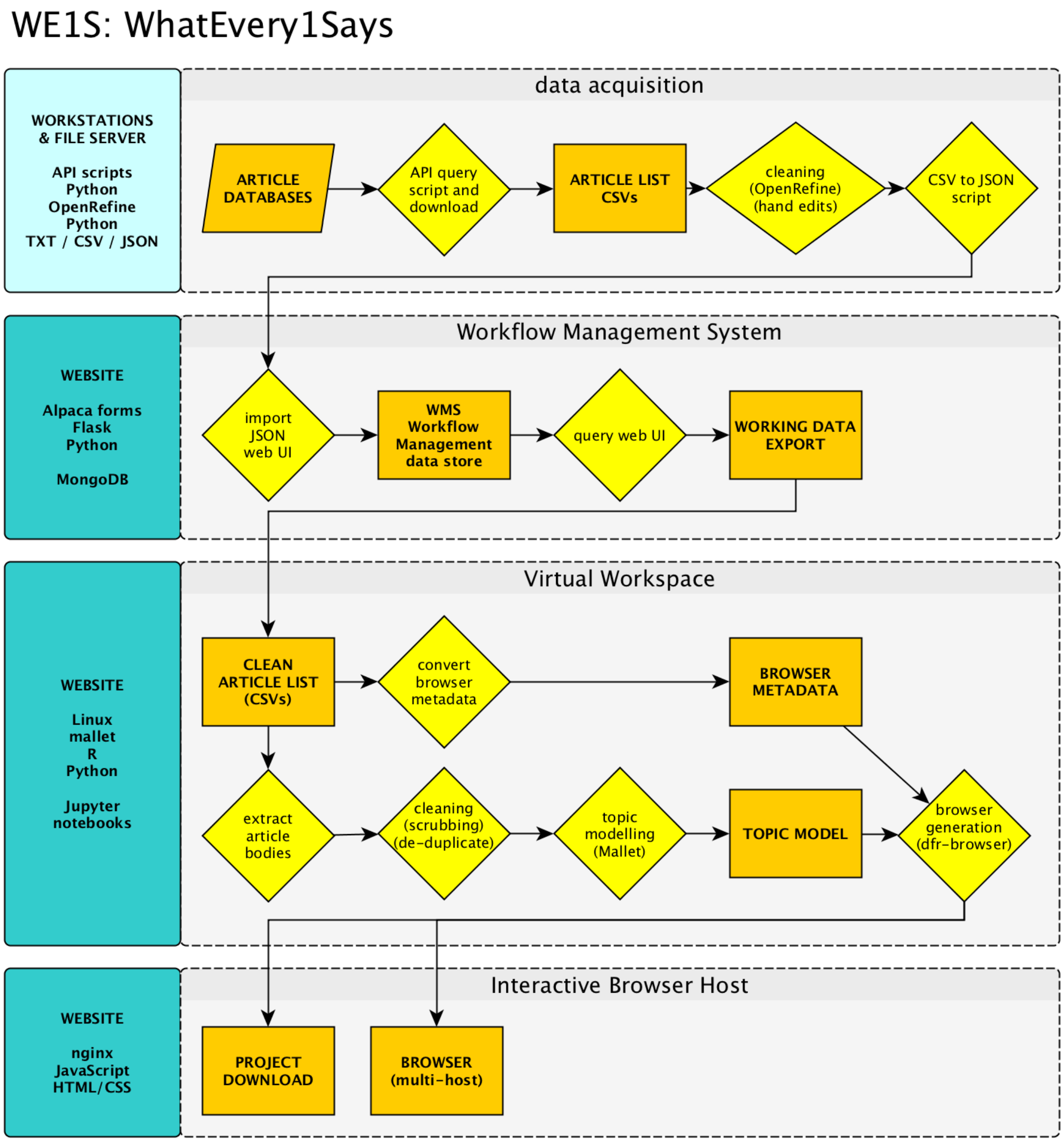


Fig. 2 Architecture of WE1S virtual workspace manager for integrated topic-modeling and visualization as implemented in a Docker virtual machine.

(4) Constructing a "Random" Comparison Corpus (Lindsay Thomas). In public discourse, there are no natural boundaries between what does and does not count as "humanities-related" discussion. The humanities, for example, can appear in both precise and general ways: as a focal topic, as part of arts and culture, in particular forms (such as literature), as part of social and ethical concerns, as part of the biographies or obituaries of individuals, etc. Indeed, it may be that one feature of the humanities is their capacity to forge multiple links between tightly focused and general themes. There is thus no pre-definable "control corpus" of public discussion on the humanities that can serve as ground truth for WE1S's topic modeling experiments. WE1S is thus using a sampled, "random" corpus as a snapshot of the larger, unclosed set of media articles to assist in exploring by contrast what articles can sensibly be defined as "humanities-specific." Doing so not only constitutes a novel approach to topic modeling in the digital humanities; it also reveals intriguing issues about the philosophy behind

statistical randomization (see Holland, 1986). This part of the panel discusses the "random" corpus WE1S created, its use within the WE1S project workflow, and includes theoretical reflections on incorporating methodology borrowed from the sciences and social sciences in DH work.

(5) "Interpretation Protocol" for Topic Models. (Ashley Champagne). One of the needs in DH research is a for way to declare not just technical but *interpretive* workflows so that they can be shared, reproduced, and evolved by the research community. In the case of DH topic model studies, for instance, rarely are there 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 human interpreter 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. WE1S has created an initial declaration of its topic-model interpretation process that defines step-by-step interactions between machine learning and human interpretation/collaboration (e.g., when in the process humans convene to interpret a topic model; what outputs, visualizations, and secondary algorithmic products such as clusterings are used; how humans discuss a topic model; how topic models and interpretive acts are iterated; etc.). The goal is to make it possible for the larger DH community to improve or vary the topic-model interpretation process in open, shareable ways.

(6) Prototyping How the WE1S Project Can Guide Humanities Advocacy (Jamal Russell). When WE1S has completed its topic models and interpretive studies, it will produce a public-facing site allowing others to explore the models and follow links to the original articles. But how can the project fulfill its ultimate ambition of guiding humanities advocacy? This final part of the panel reports on a unique early experiment in applying WE1S research. In 2016-17, a funded group of undergraduates studied sample articles from the WE1S corpus under the guidance of the project's topic models. They wrote a white paper on their findings with recommendations for humanities advocacy. And they created practical advocacy projects based on those recommendations. Using this concrete example as a springboard, the panel concludes by reflecting on the relationship between interpreting topic models and creating publicly accessible narratives about the humanities.

Works Cited

4Humanities: Advocating for the Humanities. Home page, n. d. http://4humanities.org.

______. "'What Every One Says About the Humanities' Research Project (WhatEvery1Says)." 25 April 2013. http://4humanities.org/2013/04/what-everyone-says-about-the-humanities-research-project.

Apache Taverna (Taverna Workflow System). Home page, n. d. https://taverna.incubator.apache.org.

Bare, C. (2014). "Guide to Open Science." Digithead's Lab Notebook, 9 January 2014. http://digitheadslabnotebook.blogspot.co.uk/2014/01/guide-to-open-science.html.

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

Bollen, K., J. T. Cacioppo, R. M. Kaplan, J. A. Krosnick, J. I. Olds. "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science -- Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences." National Science Foundation, 2015. http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.

Bose, R., and J. Frew. (2005). Lineage Retrieval for Scientific Data Processing: A Survey. *ACM Computing Surveys* 37.1: 1–28. https://pdfs.semanticscholar.org/3a05/2feb019328487068c8efc4c5dced8eb51a87.pdf.

DH Box. Home Page, n. d. CUNY Graduate Center. http://dhbox.org.

Garijo, D., P. Alper, K. Belhajjamey, O. Corcho, Y. Gil, and C. Goble. "Common Motifs in Scientific Workflows: An Empirical Analysis." 2012 IEEE 8th International Conference on E-Science, 2012: 1–8. DOI: 10.1109/eScience.2012.6404427.

Gil, Y., et al. (2007). "Examining the Challenges of Scientific Workflows." *IEEE Computer*, 40.12: 24-32. https://pdfs.semanticscholar.org/e45d/4aedd10229cbbeef8b2ec009f87ae1a4065e.pdf.

Goldstone, A. dfr-browser, v. v0.8a (June 8, 2016). Home page, n. d. http://agoldst.github.io/dfr-browser/.

Holland, P. W. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81.396 (1986): 945-960. http://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354.

iPython Notebook. (See Jupyter Notebook.)

Jupyter Notebook (formerly iPython Notebook). Home page, 5 March 2017. http://jupyter.org/.

Kepler Project. Home page, n. d. https://kepler-project.org.

Kleinman, S., LeBlanc, M.D., Drout, M. and Zhang, C. Lexos. v3.0. 2016. https://github.com/WheatonCS/Lexos/. doi:10.5281/zenodo.56751.

Lexos. (See Kleinman et al.)

Mallet. (See McCallum, A. K.)

McCallum, A. K. "MALLET: A Machine Learning for Language Toolkit." 2002/2016. http://mallet.cs.umass.edu/.

ProvONE. "ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance -- Unofficial draft." 27 March 2014. http://vcvcomputing.com/provone/provone.html.

WINGS (Semantic Workflow System). Home page, n. d. http://www.wings-workflows.org.