IJDC | *General Article*

Tiny Data: Building a Community of Practice around Humanities Datasets

Veronica Ikeshoji-Orlati
Vanderbilt University

Mary Anne Caton
Vanderbilt University

Suellen Stringer-Hye
Vanderbilt University

Abstract

Quantitative data, the foundation of scientific research, have been in the foreground of discussions about data creation, curation, and publication pipelines. However, data for humanistic and social scientific inquiries take many forms, including physical and ephemeral primary resources (books, objects, performances, interactions); qualitative, free-form observations; as well as quantitative, structured data and metadata. At the Vanderbilt University Jean and Alexander Heard Library, we started the Tiny Data Working Group (TDWG) in 2016 to tackle some of the humanistic research data creation and curation issues in a constructive, collaborative, and interdisciplinary format. The present paper considers what it means to be FAIR with humanities data, as well as how to build a community of data-literate humanists, based on our experiences with the TDWG.

Received 22 January 2018 ~ *Accepted* 22 January 2019

Correspondence should be addressed to Veronica Ikeshoji-Orlati, Digital Scholarship, Vanderbilt University Libraries, 230 Appleton Place PMB 135, Nashville TN 37203. Email: v.ikeshoji-orlati@gmail.com

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v13i1.663
DOI: 10.2218/ijdc.v13i1.663

Introduction

As data curators strive to make data management an integral component of the research lifecycle, it is necessary to delve into discipline-specific ways of defining and interacting with data. Contemporary principles of good data management and stewardship derive primarily from scientific research workflows, as reflected in the science-oriented language and examples presented in the recent statement of the FAIR principles (Wilkinson et al., 2016). As a result, data curation best practices are structured around the needs of scientific inquiry: facilitating reproducible research, making data machine-readable and computationally-actionable, and encouraging reuse of data in ‘downstream studies’ (Wilkinson et al., 2016). Furthermore, the growing infrastructure to sustain FAIR data, including workflow management platforms such as the Open Science Framework (OSF) and data repositories such as Dataverse, Figshare, and Zenodo, is being shaped by the needs of scientists working in collaborative laboratories on (primarily) quantitative data.

The Tiny Data Working Group (TDWG) at Vanderbilt University was convened in Spring 2017 to explore the curation and management needs of an often-neglected data creation and curation community: humanists. We started from a broad definition of data, encompassing physical and ephemeral primary sources (books, objects, performances, interactions); qualitative, free-form observations; and quantitative, structured data and metadata, too. In addition, we sought to orient working group meetings towards grappling with, and finding workable solutions to, participants’ current research questions. While we seek to align with FAIR principles in the creation and curation of humanities datasets, we have come to acknowledge that bespoke data sometimes are necessary to answer the fuzzy, open-ended questions which define humanistic inquiry. In this paper, we describe how we are building and nurturing a cohort of data-literate humanists across campus through a weekly working group format. In addition, we discuss our vision for FAIR data in the humanities and the path we are pursuing to enact it.

Data-Driven Humanities at Vanderbilt

To understand the role of the Tiny Data Working Group on the Vanderbilt campus, it is necessary to review some of the features of the campus research community. Vanderbilt is a medium-sized American research university with slightly over 2,000 graduate (primarily PhD) students and approximately 550 postdoctoral fellows (primarily in quantitative, medical, and scientific fields).[1] Of the c. 4,300 university faculty, 574 form the College of Arts and Science.[2] Digital scholarship and data science programs are at varying levels of development, but recent administrative-level efforts (including a data science visions working group[3], the establishment of the Wond’ry[4] and the Center for

[1] Graduate student and postdoctoral fellow numbers from: https://gradschool.vanderbilt.edu/about/community.php

[2] Faculty count from: https://www.vanderbilt.edu/about/facts/

[3] Vanderbilt Data Science Visions Working Group: https://www.vanderbilt.edu/provost/committees/datasciencevisions.php

[4] Wond’ry: https://www.vanderbilt.edu/thewondry/

Digital Humanities[5], and the continued support of the Vanderbilt Institute for Digital Learning[6]) have generated a critical mass of energy for envisioning new digital projects on campus.

In particular, the founding of the Vanderbilt Center for Digital Humanities in Fall 2016 has served as a catalyst for cultivating digital scholarship enthusiasm on campus by providing graduate, postdoctoral, and faculty fellowships; hosting theoretically-focused as well as practice-oriented reading and working groups; and creating a dedicated space for those curious about digital humanities to attend lectures, drop in for project consultations, and meet other faculty, staff, and students with similar interests and an array of digital skillsets. At the same time, the consolidation of the campus digital humanities community has made apparent the need for additional resources and, more urgently, specialist skills and support to make attainable the myriad digital project visions which have been imagined.

The Tiny Data Working Group: 2017-Present

In response to the growing desire and need to grapple with all aspects of the digital humanities project lifecycle, the authors (members of the Library's Digital Scholarship team) founded the Tiny Data Working Group (TDWG) in Spring 2017. In the TDWG, we work with students, postdocs, faculty, and librarians to walk through the process of crafting a data-driven humanities research project. We focus on the humanities (and social sciences) by guiding participants towards nuanced, data-driven methods for answering humanistic research questions; by demonstrating and facilitating sound data collection and management practices; and by identifying appropriate places to deposit and access data from completed projects.

'Tiny' Data: What's in a Name?

The term 'Tiny Data' was selected for two reasons: to invite traditional and digital-curious humanities scholars to the data-scholarship table and to challenge the discourse of scale as the defining feature of meaningful, data-driven scholarship. From sustained interaction with the humanities and social science research community on campus, we perceived a general reluctance to engage with broader campus data science initiatives due to the belief that their datasets were too small or could not be subjected to the same types of analysis as large, quantitative datasets. As a result, these researchers were not seeking expert guidance in developing their data-driven research methods and managing their data throughout the research lifecycle. Instead of trying to adapt humanities datasets to big data tools and methodologies, therefore, we consciously decided to work in the opposite direction and focus on how best to answer researchers' questions and workflow problems with available data wrangling and analysis tools.

Use of the term 'Tiny Data' may seem alienating to some humanists and social scientists who already consider their research to be data-driven. Nevertheless, we feel that the term embodies a critical dichotomy which the future of digital humanities must address explicitly: where and how do computational models of data analysis intersect with traditional methodologies and interpretative frameworks? The data participants

[5] Center for Digital Humanities: https://www.vanderbilt.edu/digitalhumanities/

[6] Vanderbilt Institute for Digital Learning: https://www.vanderbilt.edu/vidl/

have brought to the TDWG range from petite, manually-collected corpora of a few thousand words to unwieldy spreadsheets with dozens of columns and thousands of rows. On the whole, the data we work with in the TDWG could very well fit within the definition of 'small data' proposed by Kitchin and Lauriault: they are of small to moderate volume, often difficult to scale outside of the scope of the project, and organized around a single or a few research questions (2014). Perhaps more apropos to the research question-centric approach we have taken in the TDWG is Borgman's definition of 'little data':

> 'Data are big or little in terms of what can be done with them, what insights they can reveal, and the scale of analysis required relative to the phenomenon of interest' (Borgman, 2015).

We intend to maintain the term 'Tiny Data' for our working group, however, because the phrase carries additional connotations. Outside of scholarly publications, the term 'tiny data' has been applied to data that are challenging because they are incomplete or seemingly insufficient.[7] The phrase is also used to describe individual interactions and moments of engagement in customer service.[8] Fundamentally, we believe that 'Tiny Data' are at the core of traditional and contemporary humanistic inquiry, reflecting scholars' critical engagements with texts, images, sound, and performance. As research collaborators and facilitators, we seek to equip humanists and social scientists with the necessary data curation skills to enable them to explore and answer their research questions with whatever combination of traditional and computation methods they see fit, then share their data and results with one another in a public, well-documented, and easily-accessible manner.

Spring 2017: Working Group Formation

The practical origins of the TDWG lie in the Fall 2016 THATCamp hosted by the Center for Digital Humanities at Vanderbilt University. As indicated above, the digital humanities community was beginning to coalesce around the Center for Digital Humanities, supported by the Digital Scholarship team in the Library. A THATCamp session on the use of graph databases in humanities research, led by Suellen Stringer-Hye, generated extensive debate on the challenges and opportunities presented by using digital tools to analyse analogue materials. In particular, several graduate students expressed reluctance to move away from the close readings which characterized their methodologies, as well as wariness of employing sterile, Big Data approaches to humanistic research more broadly. The authors had already started work on defining 'Tiny Data' and the relationship between it and computational approaches to humanistic research; when the graduate students expressed interested in experimenting with their research topics through the controlled application of data management and visualization tools, the first iteration of the TDWG was convened.

In Spring 2017, a handful of graduate students from the French and History departments, alongside librarians with archaeological and curatorial interests, gathered weekly to discuss various facets of data curation and visualization for the humanities. Topics included how to collect high-quality digital data during an archival visit, how to

[7] C.f. an exercise to visualize a tiny dataset of two numbers: https://www.thedataschool.co.uk/alexandra-hanna/tiny-data-in-tableau/

[8] Forget Big Data: How Tiny Data Drives Customer Happiness: https://blog.trello.com/forget-big-data-how-tiny-data-drives-customer-happiness

identify the data points necessary to answer specific research questions, and a survey of digital tools and data models which could help elucidate the methodological strengths and weaknesses of participants' research questions. Each week, one or two participants would present the data or research question they were working on and the group would collaboratively workshop the data or idea.

A sustained interest in data visualization, and particularly in network visualization, lead us to continue the THATCamp discussion of graph databases throughout the semester. Some participants even elected to import sample datasets into neo4j to determine whether a network visualization would help them answer their research questions and, if so, how much additional data would be required to create a useful model. The deliverables from the first semester of the TDWG included not only the completion/initiation of two research projects, but also something less tangible: the formation of a cohort of humanities graduate students with a heightened understanding of the strengths and pitfalls of digital data-modelling tools and a new perception of the role the Library could play in the formative stages of their research.

Fall 2017: Establishing a Curriculum

During the Fall 2017 semester, the Tiny Data Working Group grew to include graduate students, postdocs, faculty, and librarians from across the humanities and social sciences. Members of the Center for Digital Humanities, the Visual Resources Center, the Provost's Office, and the Departments of Anthropology, German, History, Psychology, Russian and East European Studies, and Music joined. In response to questions about systematic data collection, curation, and preservation protocols which had arisen in the preceding semester, we shifted the working group meeting structure towards a blended discussion/workshop model.

As is illustrated by the working group syllabus, discussion sessions were based around selected readings and critical analyses of digital humanities projects and their data.[9] In particular, we discussed both the hard and soft skills necessary to build, sustain, and sunset a data-driven digital humanities project. The goal of the discussion sessions was three-fold: to sketch out a roadmap/workflow for humanities data curation through the selection and sequencing of relevant topics; to identify successful models for humanistic data collection and sharing practices; and to cultivate a collegial and collaborative atmosphere amongst participants. The 'Bring Your Research (Data)' Workshops were presented as opportunities for hands-on guidance with tools for data cleaning, modelling, and publication. During the workshop sessions, we introduced OpenRefine, a selection of metadata schemata, the concept of controlled vocabularies and the Linked Open Vocabularies site[10], and subject-specific vs. content-agnostic data repositories.

The Fall 2017 TDWG participants each worked towards gathering and curating a dataset based on their individual research questions. As part of the process, we encouraged participants to generate thorough and thoughtful metadata to describe the data they were collecting, as well as long-form explications of their data collection and curation methodologies regardless of how complete (or not) their projects were. By leveraging the diversity of experiences and perspectives represented in our Fall 2017 TDWG cohort, we were able to iteratively improve participants' data collection workflows, data standardization practices, and plain-language documentation during the

[9] The syllabus for the semester is available here: https://github.com/HeardLibrary/tiny-data/blob/gh-pages/Fall-2017/syllabus.md

[10] Linked Open Vocabularies: http://lov.okfn.org/dataset/lov/

workshop sessions. Moreover, by encouraging active discussion amongst participants and facilitating hands-on workshops utilizing participants' own data, we were able to model a version of solo humanistic research which was not solitary.

Spring 2018: Collaborative Data Wrangling

As the Spring 2018 term gets underway, we are adapting the format of the TDWG again in response to participants' feedback. In particular, Fall 2017 participants wanted to have more hands-on experience with specific data curation tools, such as OpenRefine. Furthermore, they expressed a desire to receive more exposure to, and guidance for, utilizing other digital tools, structures, and languages (such as the NLTK and MySQL) for streamlining their research and data curation workflows. To ensure that participants' projects are moving forward throughout the semester, each one has uploaded their data, in whatever form, to a Box account shared with all group members so that each week, participants can follow along as we clean, model, and visualize a participant's data.

In addition, at the request of Fall 2017 participants, we are working towards creating documentation for how to start and sustain a humanistic, data-oriented project. Whereas the Digital Humanities Data Curation Guide (Flanders and Muñoz, n.d.) provides an excellent model for grappling with the big-picture issues of creating FAIR humanities data, we hope to produce a more pragmatically-minded resource to help other humanists think through their data and how to manage and curate it for the future.

Take-Aways

Over the past year, the TDWG has grown and flourished in ways we had not expected when it was first convened. While the TDWG is an ongoing project, we propose it as a useful and replicable model for reaching humanists in a research data management program. A few of the things we have learned over the past year are outlined below.

Rethinking FAIRness for Humanities Data

The success of the TDWG is due, in part, to the fact that it serves as an entry point for diverse types of humanities and social science researchers to make data curation part of their research workflows. We have accomplished this by keeping the focus on human-scaled datasets and embracing traditional, manual methods for data collection and curation in our working group dialogues. The frequent inconsistencies and incompleteness of humanities data, however, alongside many scholars' desire to represent qualitative observations of, and interactions with, primary sources in their datasets, make it challenging to standardize all humanities datasets into fully FAIR-compliant objects.

While we do not advocate for 'messy' data in the TDWG, we contend that high-resolution humanities research questions sometimes require more nuance than rigid adherence to standardised vocabularies and schemata may allow. As a result, the reusability and interoperability of the data may be limited, since bespoke data frequently have little use outside of their original context. Furthermore, the importance of making data fully interoperable and reusable rarely resonates with humanists who are accustomed to collecting and analysing 'tiny' datasets to answer a particular research question.

Nevertheless, it is critical to introduce and reinforce good data curation practices amongst humanities scholars, regardless of whether their preferred research methodology is computational or analogue or whether their data are a few hundred images or a gargantuan spreadsheet. Translating abstract data management concerns such as findability, accessibility, and reusability into pragmatic issues such as how to organize data during an archive visit and how to normalize data so that a desired visualization can be made has been key to engaging and sustaining a robust cohort of TDWG participants in a longer-term conversation about data literacy. By introducing metadata documentation standards in the context of facilitating future development of a research project and discussing data publication as a citable, potentially peer-reviewed publication, we have moved even the most reticent of digital humanists towards creating FAIR(er) data.

From Research Questions to Data Curation Methods

Intertwined with the reconsideration of what it means to be FAIR with humanities data is the importance of privileging TDWG participants' research questions over specific data curation methods and tools. Since its inception, the TDWG has been targeted towards both more and less traditional humanities scholars, the former of whom are often alienated by the scale and methods of digital humanities research. By focusing on common ground – an actual research question and how to make the discovery and analysis process go more smoothly – we are succeeding in cultivating a diverse cohort of data-literate humanists across campus. We consider this our nascent community of practice for humanities data curation and are eager to facilitate turning the TDWG participants into data curation evangelists amongst their colleagues.

Redefining the Role of the Library

The TDWG was formed during a pivotal moment in the landscape of digital scholarship on the Vanderbilt University campus. While energy and enthusiasm for creating digital projects was (and remains) on the rise, the systematic infrastructural support to generate and sustain data-driven digital humanities projects has been falling behind. Indeed, the growth of the TDWG cohort reflects a need for more data stewardship guidance at the naissance of projects.

As the paradigms of digital, data-driven scholarship continue to shift in the coming years, the TDWG has created a new, organic pathway for positioning librarians as collaborators throughout the research process. Librarians participating in the TDWG, for example, are stepping up to act as way-finders and translators between research questions and the methodologically-sound application of new data wrangling and analysis tools. The success of the TDWG in redefining the role of the librarian in generating innovative, data-driven humanities scholarship as well. Recently, there has been a growing number of requests for librarians to serve as consultants on data-driven initiatives across campus, as well as at special events such as the Bring Your Data workshops, initiated by the Center for Digital Humanities and inspired by the Fall 2017 TDWG.

Next Steps

Over the next semester, the TDWG will continue to train the current cohort of participants in how to integrate good data management and curation practices into their research workflows. By producing concrete deliverables in the form of publishing well-formed and documented datasets, we hope to grow the community of humanities data curation practitioners in the TDWG.

In addition, we will focus on generating documentation of the pathways humanities scholars may take to go from a research question to a fully-fledged, data-driven research project with data that is as FAIR-complaint as possible. As part of that project, we are collaborating with our Scholarly Communications Librarian, Elisabeth Shook, to identify ways to develop humanities data collections in the Institutional Repository, DiscoverArchive.

Finally, on a more theoretical note, we aim to rethink what it means to have FAIR humanities data. Is it possible to generate FAIR humanities data by building more communities of practice around the issues surrounding the diverse types of research question scales and data types which define humanistic inquiry in the 21st century?

References

Borgman, C.L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3.* doi:10.1038/sdata.2016.18

Flanders, J. & Muñoz, T. (n.d.). An introduction to humanities data curation. Digital Humanities Data Curation Guide. Retrieved from http://guide.dhcuration.org/contents/intro/

Kitchin, R. & Lauriault, T.P. (2014). Small data in the era of big data. *GeoJournal 80(4)*, 463-475. Retrieved from https://link.springer.com/article/10.1007%2Fs10708-014-9601-7