Liam Downs-Tepper
Digital Humanities Praedoc
University of Vienna Digital Humanities

Rotten References, Dead Data, and Lost Links: A Macro/Micro Examination of Sustainable Data

I have data files from projects from years ago which are on disks I no longer have a drive for on computers I no longer have access to or are no longer made…[1]

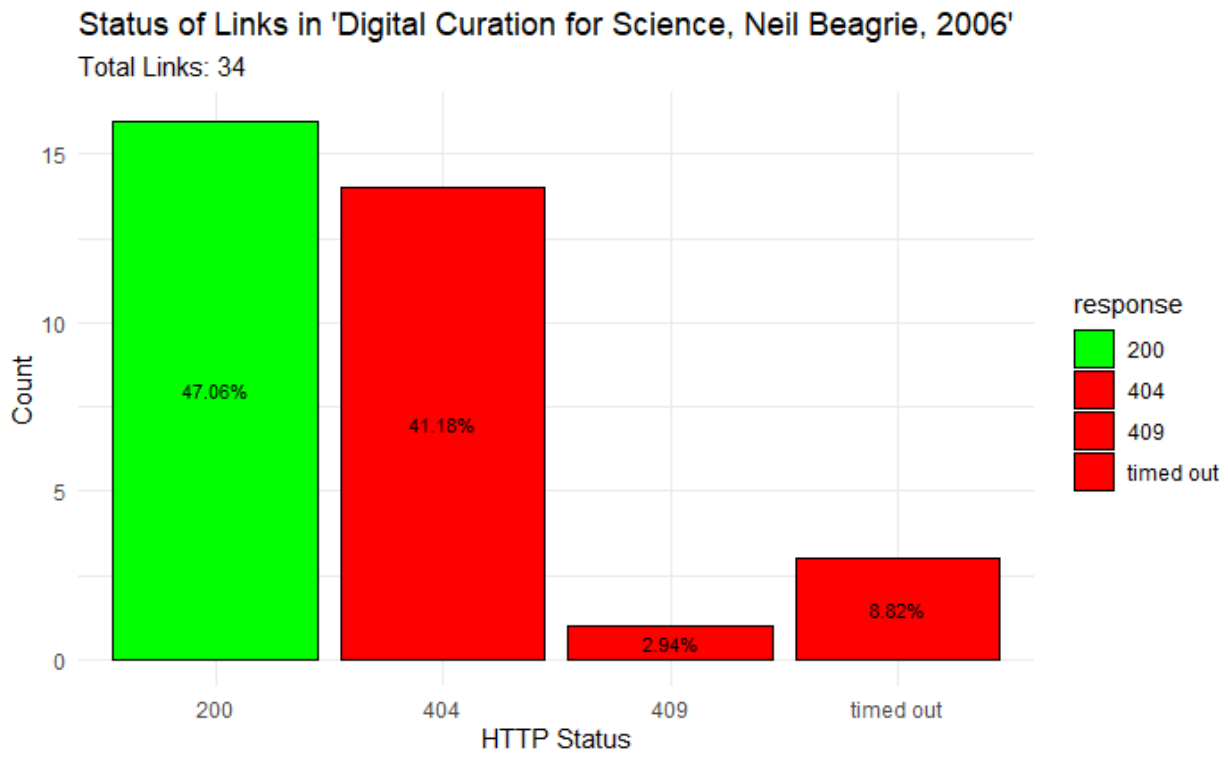
Although the rationale behind it remains murky, it can be generally agreed that thirteen has long held the unenviable title of "Unluckiest Number" for much of the globe.[2] In this regard, superstition and data overlap: numbers hold far more meaning than they would in other contexts. Good data, accessible information, crunchable numbers… these are our bread and butter. But while a seasoned researcher would not flinch from walking past a black cat, under a ladder, on the thirteenth floor, they will always recoil in horror from that most feared symbol: Error 404.

Handling and preserving of data is an eternal discussion in the digital humanities. One can find tutorials from the 1970s on handling data as a humanist, which even then recognized the lack of common vocabulary that exists between computer scientists and humanists.[3] There are countless similar guides and tutorials spanning the 53 years since this was first published, posing related questions: How should data and sources be handled?[4] How do we make data sustainable?[5] These tutorials and articles offer a slew of standards, a tsunami of tools, a deluge of databases. Always something new: "Digital Humanities is a production-based endeavor…"[6]

But what happens to everything that we, as academics, produce? A 2006 paper on data curation, "concludes that significant effort needs to be put into developing a persistent information infrastructure for digital materials and into developing the digital curation skills of researchers and information professionals."[7] Such infrastructure is a necessity for creating work that is reproducible and verifiable.

---

[1] Interview included in: Lord and Macdonald, "The JISC Committee for the Support of Research (JCSR)," 17.
[2] Markovsky, "Why Is 13 Considered Unlucky?"
[3] Sedelow, "The Computer in the Humanities and Fine Arts," 89.
[4] Caplan, "Method without Methodology: Data and the Digital Humanities."
[5] Poole and Garwood, "Digging into Data Management in Public‐funded, International Research in Digital Humanities"; Kinnaman and Guimont, "DH as Data."
[6] Burdick, *Digital Humanities*, 13.
[7] Beagrie, "Digital Curation for Science, Digital Libraries, and Individuals," 3.

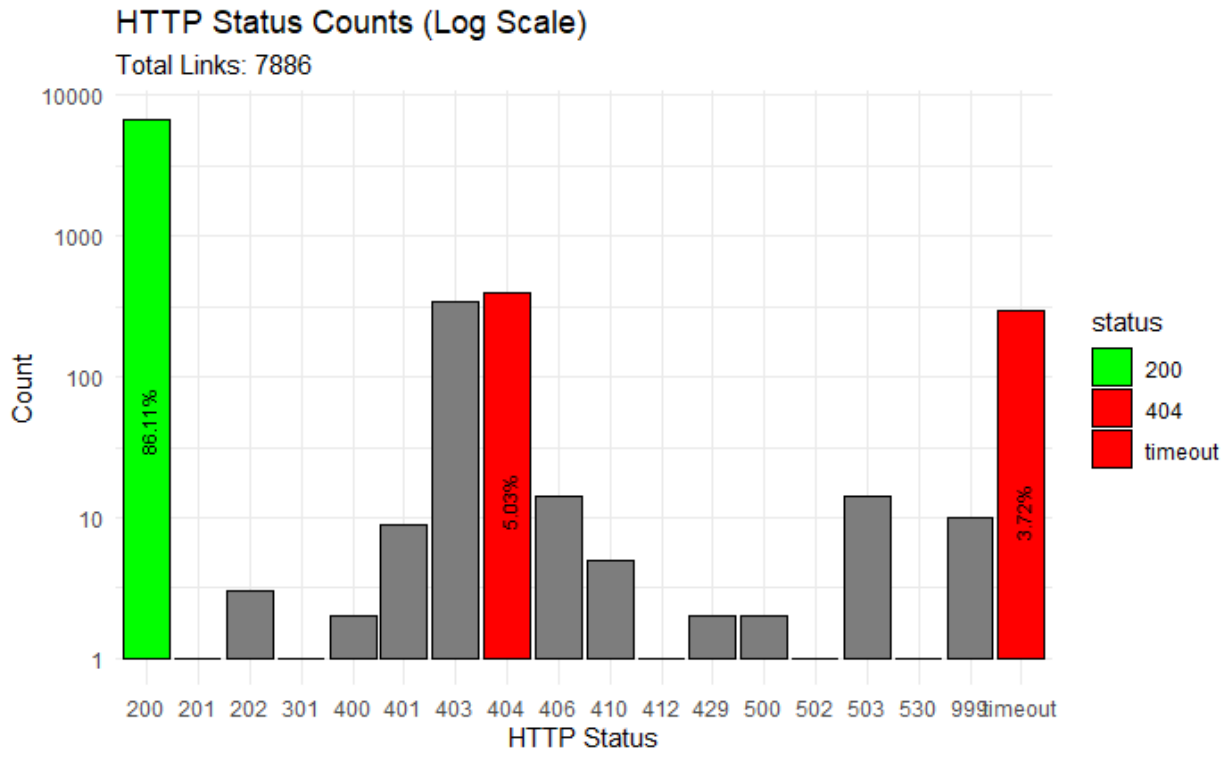


Unfortunately, the infrastructure is not there yet; the above chart offers (in vivid technicolor!) a look at what is more likely to happen: more than half of the links included in that same article on data curation no longer work.[8]

This paper aims to explore the degradation and disappearance of data and scholarship in the digital humanities through a hybrid macroscopic/microscopic approach. To assess link and reference rot broadly, this research opts for embracing automation to aid with URL testing. Two starting sources were chosen, each offering a wealth of links to data, tools, and publications. First, *Data is Plural,* a weekly newsletter of interesting datasets which has been in circulation since 2015.[9] Over the course of more than 350 editions one can find nearly 8000 links to any and all things data related. Second, *Digital Humanities Quarterly,* an online DH publication available since 2007. DHQ's articles contain in total over 13,000 links, and are available as openly accessible XMLs. Expanding this analysis, a number of other open access DH and DH-adjacent publications have been examined as well, including: ARIADNE, Humanites Numeriques, the Journal of the Text Encoding Initiative, Zeitschrift für digitale Geisteswissenschaften, Current Research in Digital History, Digital Medievalist, Le champ numerique, Interdisciplinary Digital Engagement in Arts & Humanities (IDEAH), Internet Archaeology, The Journal of e-Media Studies, the Journal of Electronic Publishing, and the Journal of Interactive Technology and Pedagogy.

[8] Links scraped from: Beagrie, "Digital Curation for Science, Digital Libraries, and Individuals."
[9] Singer-Vine, "Data Is Plural."

The combination of these different sources can help offer further insights into different approaches to data, data journalism, and archiving. Initial research has already proved fruitful; automated testing has been shown to be an effective way to get a broad view of the situation. Exploratory data analysis of HTTP returns shows that more than 86% of *Data is Plural* links continue to work without any issues (see chart below), in contrast to only 76% of DHQ links. While HTTP status is an imperfect metric for assessing availability, it is one of the best for operating at scale; across ARIADNE's 79 issues, one can find 18,000 links. This builds off of the technical methodology outlined by Klein et al.[10]



The second part of this research involves contacting the creators of now-defunct data, authors of particularly disconnected articles, and hosts of abandoned links, to inquire what happened to the projects in question. This is done through semi-structured interviews, aiming to highlight which structural issues are obstacles, or if issues are simply a question of planning gone awry, or something else entirely. Semi-structured interviews are particularly useful for this type of work, as they allow for greater comparability without restricting, guiding, or limiting interviewee options. Initial forays into questioning to this effect have proven valuable, revealing intriguing results:

- Data migrated on museum websites from "active exhibitions" to "archived exhibitions" led to broken links for a number of authors aiming to highlight then-current cultural events

[10] Klein et al., "Scholarly Context Not Found."

- A failed archival startup created as a response to much larger archival infrastructure, as part of a nationalist push to not use European/American data tools.
- Outsourcing of data storage and preservation by an author, who was surprised that the data had vanished.
- Private individuals digitizing data without the knowledge of its source, but only leaving the source as attribution.
- A museum began as a private business which originally used the business URL, but has since moved to its own website - and abandoned the original page.

This work builds upon many before it. To highlight only a few: Drucker highlights some of the limitations that researchers must be aware of when aiming for scoping and sustainability, noting that issues can come in as "technical, institutional, intellectual, and financial."[11] Coble and Karlin explore reference rot in DHQ specifically, but take a very different angle on testing samples.[12] Da Silva and Nazarovets provide perhaps the most comprehensive overview of work on the topic, though their goal is ultimately to provide a solution: the Internet Archive.[13]

The hybrid approach offered here is distinct from other methods used thus far, and gives an effective way to get a well-rounded, multi-level view of the real state of data and link preservation among data scientists and digital humanities. Macro analysis gives both a broad overview, and is easily scalable, enabling the inclusion of other journals, data pages, and so on. The two initial data sources proposed work to give useful temporal and thematic range, sufficient to evaluate the severity of these issues over time. Macro analysis also highlights potential interviewees for micro analysis. Ultimately, this paper offers a unique approach to get a clear picture of the state of data preservation in the Digital Humanities today.

[11] Drucker, "Sustainability and Complexity."
[12] Coble and Karlin, "Reference Rot in the Digital Humanities Literature."
[13] Teixeira da Silva and Nazarovets, "Archiving Website‐based References in Academic Papers."

Works Cited

Beagrie, Neil. “Digital Curation for Science, Digital Libraries, and Individuals.” *International Journal of Digital Curation* 1 (2006): 3–16. https://doi.org/10.2218/ijdc.v1i1.2.

Burdick, Anne, ed. *Digital Humanities*. Paperback edition., 2016.

Caplan, Lindsay. “Method without Methodology: Data and the Digital Humanities,” 2016.

Coble, Zach, and Jojo Karlin. “Reference Rot in the Digital Humanities Literature: An Analysis of Citations Containing Website Links in DHQ.” *Digital Humanities Quarterly* 017, no. 1 (May 26, 2023).

Drucker, Johanna. “Sustainability and Complexity: Knowledge and Authority in the Digital Humanities.” *Digital Scholarship in the Humanities* 36, no. Supplement_2 (November 5, 2021): ii86–94. https://doi.org/10.1093/llc/fqab025.

Kinnaman, Alex, and Corinne Guimont. “DH as Data: Establishing Greater Access through Sustainability.” *Digital Humanities Quarterly* 017, no. 3 (August 24, 2023).

Klein, Martin, Herbert Van De Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. “Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot.” Edited by Judit Bar-Ilan. *PLoS ONE* 9, no. 12 (December 26, 2014): e115253. https://doi.org/10.1371/journal.pone.0115253.

Lord, Philip, and Alison Macdonald. “The JISC Committee for the Support of Research (JCSR),” 2003.

Markovsky, Barry. “Why Is 13 Considered Unlucky?” University of South Carolina. Accessed February 6, 2024. https://www.sc.edu/uofsc/posts/2022/10/conversation_thirteen.php.

Poole, Alex H., and Deborah A. Garwood. “Digging into Data Management in Public‐funded, International Research in Digital Humanities.” *Journal of the Association for Information Science and Technology* 71, no. 1 (January 2020): 84–97. https://doi.org/10.1002/asi.24213.

Sedelow, Sally Yeates. “The Computer in the Humanities and Fine Arts.” *ACM Computing Surveys* 2, no. 2 (June 1970): 89–110. https://doi.org/10.1145/356566.356568.

Singer-Vine, Jeremy. “Data Is Plural,” December 20, 2023. https://www.data-is-plural.com/.

Teixeira da Silva, Jaime, and Maryna Nazarovets. “Archiving Website‐based References in Academic Papers: Problems Caused by Reference Rot, Potential Solutions and Limitations.” *Learned Publishing* 36 (July 12, 2023): 477–87. https://doi.org/10.1002/leap.1560.