# 'Everyone has their reasons for curating the data they have decided to keep': a thematic analysis of data hoarding as digital curation practice

*Emily Maemura and Travis L. Wagner*

## Abstract

**Introduction.** This paper presents a preliminary analysis of the r/DataHoarder subreddit, an online community focused on social, legal, and technical practices related to digital curation. We contrast how they conceptualize their practices with the models, frameworks, and capabilities established for digital curation in more traditional memory institutions.

**Method.** We use thematic analysis to analyse top posts (n=170) from the subreddit, to determine how members conceptualize, describe, and enact their data hoarding practices.

**Findings.** Two key themes are identified: a focus on materiality of storage for hoarding, and use of the subreddit to promote calls-to-action. Each theme is further analysed, identifying underlying motivations such as nostalgia and vigilantism.

**Discussion.** We briefly discuss the theoretical and practical implications of these findings in contrast with workflows like the DCC curation lifecycle, while also addressing limitations of this preliminary work, and outlining potential new research directions moving forward.

**Conclusion.** We have found this analysis presents useful counterpoints to commonly referenced and standardised practices of memory institutions. We believe this will continue to be a fruitful area of study as we conduct future work on how r/DataHoarder and similar communities conceptualise, practice, and develop their own ethical framework toward sustained access of digital materials.

# Introduction

The r/DataHoarder subreddit is an online community focused on social, legal, and technical practices related to digital curation. Created in May 2013 and with a current membership of over 700,000 users, it is neither representative of any institution nor defined by a shared digital curation credo. Yet, r/DataHoarder nonetheless defines itself as a community of *'digital librarians'* who have varied *'reasons for curating the data they have decided to keep (either forever or For A Damn Long Time<sup>tm</sup>)'* (r/DataHoarder, 2017). Regardless of type of data or reasons for hoarding, r/DataHoarder members share best practices, technical guidance, and, on occasion, ethical stances towards the increasingly complex sociotechnical landscape of a digital-first world. As such, r/DataHoarder acts as a distributed community of practice, enacting digital curation through deinstitutionalized and extralegal means.

A recent example of r/DataHoarder's collective efforts is seen in their archiving work in the wake of the January 6th United States Capitol attack. The subreddit's users collaborated to document and preserve footage from the attack as well as communications between those involved, which received both scholarly and journalistic attention (Basu, 2021; Chapman, 2022). This example demonstrates how, despite a distributed community structure, r/DataHoarder's members also coalesce around shared ethics and practices. We take an interest in exploring these motivations, where they both align with professional guidelines, or alternately exhibit distrust for traditional collecting institutions (i.e., government agencies), and how they work intentionally outside these institutional boundaries.

This study therefore examines how r/DataHoarder imagines and actualizes digital curation practice. We consider how such framings align with and depart from standards and practices of digital curation within traditional memory institutions. We present a preliminary thematic analysis of the top posts which have a vote count of 2,000 or more (Panek, 2021). We present this exploratory study to address the following research questions:

- What is the practice of *'data hoarding'* and what activities and capabilities does it entail?
- Where do these activities and capabilities intersect with or diverge from those of memory institutions' practices of curation?
- What do these practices reveal of the ethics and social ideologies underlying data hoarding work, and motivations of individual data-hoarders?

# Literature review

Over the past three decades, scholarship in digital preservation and curation has produced a knowledge base of best practices, primarily focused on conducting this work within organizational settings. This includes the standardization of preservation systems and functions through the open archival information system reference model, as well as common practices and activities as referenced in the digital preservation coalition (DPC) handbook and outlined in the digital curation centre (DCC) lifecycle. The specific needs for organizational and individual skillsets in conducting preservation and curation have also been studied, prescribed, and measured through capability models and tools like the NDSA levels of preservation, DPC competency audit tool, and DLF's levels of born-digital access (NDSA, 2019; McMeekin & Currie, 2022; Peltzman et al., 2022).

Largely separate from these organizational approaches, we see increased public awareness of digital content at risk of loss. Recent grassroots collecting of digital materials include: work by the environmental data governance initiative to collect EPA data prior to the beginning of President Trump's term in office in 2017 (Walker et al., 2018); work by archive team to collect NSFW materials following Tumblr's policy change and removals in 2018 (Ogden, 2022); efforts of Saving Ukrainian Cultural Heritage Online (SUCHO) to preserve digital materials from Ukraine following the Russian invasion in February 2022 (LeBlanc et al., 2022). We observe that these forms of collecting, and

preservation differ significantly from the work by practitioners within traditional memory institutions. We take inspiration from Dallas (2016) who advises digital curation research should,

*focus on the study of actual practices of curation in a diversity of contexts ... and should prioritize such intellectual inquiry over the imposition of models developed within the custodial fold* (p. 440).

We therefore take up a concern with understanding collecting and curation outside of memory institutions.

For this initial, exploratory study, we frame this work as a '*community of practice*' to understand how groups of diverse and distributed individuals engage in digital curation. Coined by Lave and Wenger (1991), the concept of '*communities of practice*' (CoPs) originates in organizational studies. Cox (2005) addresses the concept's varied usage and distinctions between the early applications of CoPs, and later take-up by Brown and Duguid (1991, 2001), as well as Wenger's (1998, 2000) subsequent publications. In information studies, the concept is often deployed to describe or facilitate efforts in community building around shared practices, both within and across organizations. For digital curation in particular, CoPs have been the guiding framework for developing capacity-building workshops at the local and national level (Moran et al., 2019, Rios et al. 2020).

We adopt a broader view of CoPs, as advocated by Bowker and Star (2000) who place less emphasis on the organizational setting. This understanding of CoPs has been previously applied to study the knowledge practices of amateur and hobby communities, and particularly their distributed formation in virtual space (Hills, 2015). Several precedents also explore this type of community formation for learning and sharing practices specifically through reddit and its structure of subreddits (Britt et al., 2022; Haythornthwaite et al., 2018; Hudgins et al., 2020; Kwon et al. 2020). Proferes et al. (2021) provide a systematic overview of past research with reddit, which we also use to inform our choice in data collection, scale and methods of analysis, and ethical decisions for working with public but potentially sensitive and personal posts and comments.

## Method

This preliminary study deploys thematic analysis to examine how members of the r/DataHoarder subreddit conceptualize, describe, and enact their data hoarding practices. Thematic analysis allows researchers to qualitatively surface ideas generated by individuals, communities, or institutions. As noted in Braun and Clarke (2006), theme identification does not necessarily correlate with a quantitative measure such as prevalence in the dataset. We follow Braun and Clark's six phases of thematic analysis: familiarization with data; generating initial codes; searching for themes; reviewing themes; defining themes; and reporting findings (Braun & Clarke, 2006, p. 87-93). We employ thematic analysis alongside theoretical commitments from CoPs research, whose concepts serve as guide points for framing themes deductively, while still offering space for inductive themes to emerge.

Reddit has been studied using thematic analysis within scholarly fields ranging from gender and sexuality studies to addiction medicine (Proferes et al., 2021; Graves et al., 2022; Graham & Rodriguez, 2021). When applying thematic analysis towards Reddit as a site with multiple CoPs, we note past work that addresses the complex, often socially contentious practices this platform affords users, especially as a means to examine topics which they might face criticism or institutionalized intervention (Lundeen et al., 2024; Maxwell et al., 2020).

### Data collection

We began data collection by gathering the subreddit's most popular posts throughout its history (as of August 2024). A cutoff point of 2000 votes (tallying upvotes and downvotes) was chosen for this initial analysis, resulting in 170 total posts included in the analysis. Posts include a range of

text-based original posts, cross-posting of links to outside news articles, photographs, and memes. The links to these 170 posts were documented within a shared spreadsheet for purposes of analysis. Post titles, authorship (reddit username) and select excerpted comments were also collected for select posts but will not be shared through direct quotes or reproduced in the analysis here. While this information is available on the public web, we follow guidance from Proferes et al. (2021) to respect the privacy of social media users, especially as this initial analysis is not yet able to address the extent of sensitive and personal information shared within these posts or the potential for users' identities to be triangulated through other posts and comments made on reddit.

## Data analysis

Following the iterative process described in Braun and Clarke (2006), the researchers deployed constant comparative coding to identify both emergent themes, while also working to create shared definitions for the themes which emerged. Work began with coding the top ten percent of the total posts within the dataset (n=17 posts), ranked by vote-count. For each post, we independently analysed the text, images, comments, and interactions to develop an initial set of codes that would inform the development of overarching themes. Upon identifying these codes, we then developed an initial set of themes that applied to these first 17 posts, before proceeding to validate these themes against the subsequent top posts to ensure that no other salient codes emerged. As an intermediate step in this iterative process, a set of high-level codes emerged. Sample codes included: materiality; value; and preventing loss. We then reviewed the remaining posts, each focusing primarily on identifying posts and comments aligned with one of these themes. A final round of discussion and refinement led to the identification of two key, overarching themes which we characterize and discuss in greater detail below.

# Findings

Based on this initial analysis, we identify a set of key themes that define the r/DataHoarder vision of digital curation as a practice.

## Theme (1) materiality of hoarding

Many of the top posts in r/DataHoarder highlight the materiality of hoarding, through imagery of storage media (e.g., photos of hard drives) as well as discussion of services and costs (e.g., fees for cloud storage or high-bandwidth ISP usage). Within this space, talking about and demonstrating the capacity for data storage becomes a central signifier of one's membership in the community. Practices enacted around materiality include sharing information on acquiring storage (notices of hard drives on sale), as well as sharing advice about in-home storage set-ups. We also take materiality to include the processes of acquiring data or digital objects, such as offline downloads of '*all of Wikipedia*,' and comments often signaling the creation of such a collection as a telltale sign of digital hoarding. Materiality as presented by r/DataHoarder posts often conflates data and digital objects, and ultimately emphasizes having control over digital materials through such storage.

### Subtheme 1.1 fetishization and bragging

An additional facet of this focus on materiality emerges in the fetishization of cutting-edge hardware and technologies such as multiple petabyte hard drives. One post garnering significant comments and discussion includes a photo of the hard drives storing the data used to generate the first ever photographic image of a blackhole in 2019. Indicative of the fixation not on the black hole photograph itself, but instead of the storage devices, commenters metaphorically fetishize the hardware via suggestions of sexual pleasure in response to seeing these images. Another popular post presents the message received from a cloud hosting platform, that the user has exceeded their storage capacity, representative of a trend in which members brag about surpassing limitations on data plans. Not only are members concerned with storage capabilities, but the

aspect bragging about capacities of hardware or cloud-based storage, often fetishizing the ability to store previously unimaginable amounts of information.

### Subtheme 1.2 nostalgia

In contrast to the fetishization of ever-newer, bigger storage, we observe how nostalgic leanings also shape the view of digital materials. One post describes a '*collapse-proof laptop,*' with comments discussing the utility of smaller, often obsolete formats and storage hardware as a mode to counteract anticipated corporate and institutional failures harkening back to a '*simpler time*' of computing's past. This aligns with a survivalist mentality of '*internet collapse preppers*', (r/DataHoarder, 2017) and the need to create tools which function without pervasive online connections and cloud services (Grandhi et al., 2020). In the same post, members suggest retaining copies of video games such as Tetris and Quake II which reflect low data demand yet represent high yield entertainment. The role of software from one's youth and a tendency to focus on collections around retro-gaming is also indicative of this nostalgic turn, as seen in another post referencing collecting rare data tapes from Nintendo 64's development. The interest in materiality therefore often overlaps with this nostalgia for obsolete hardware and formats which recalled simpler ecosystems for hoarding data and digital objects.

## Theme (2) call-to-action

The call-to-action is one of the themes we largely anticipated at the outset of this study, as it aligns with r/DataHoarder activities related to the January 6th United States Capitol. Since the subreddit's description focuses largely on its role as a space for '*like-minded individuals to exchange strategies, war stories, and cautionary tales of failures,*' (r/DataHoarder, 2017) these moments that we identify as calls-to-action are rather exceptional in relation to the day-to-day discussions on the subreddit. Importantly, these interventions provide the rare opportunity for data-hoarders to directly collaboration and take collective action. Many calls-to-action revolve around urgent collecting of materials directly at risk due to corporate decisions for removal, as well as political turmoil or threats of censorship. For instance, an additional example from our dataset calls on the r/DataHoarder community to help 'backup' the social media and website of Hong Kong Stand News after their headquarters were raided and staff arrested by national security police in 2021.

### Subtheme 2.1 vigilance

Related to direct calls-to-action, we observe that participants in r/DataHoarder express the need to remain vigilant, or '*stand watch*' in the face of potential future threats. This includes activities such as: sharing details on proposed changes to legislation affecting data transfer, privacy, and encryption; news of legal cases won or lost by perceived adversaries (namely large companies like Apple, Microsoft, YouTube, Yahoo, as well as Reddit itself), as well as ongoing legal issues experienced by the perceived ally, non-profit Internet Archive.

### Subtheme 2.2 vigilantism

Aligned with the overarching theme of call-to-action, the motivations driving many individuals to act often appear related to an ideal of vigilante justice. This surfaces in direct ways, as r/DataHoarder posters position their work as countering the aims of unethical and unjust companies. The r/DataHoarder community therefore justifies the use of hacking, piracy, and extralegal means for taking data out of the hands of these corporations, and hoarding is perceived as contributing to the public good of making information free, open, and accessible. Several examples relate to specific projects in open science and open culture, including one post linking a repository of thousands of scientific articles on COVID-19, made openly available through torrents and decentralized web links, referencing deceased hacktivist Aaron Swartz. Beyond these direct references to piracy and hacktivism, the vigilante ethos arises across the subreddit's discussions, for instance a post offering a 'bounty' for recovery of a digitized episode of a prominent 1980s TV talk show.

## Discussion

The analysis of these themes reveals where the activities and capabilities of data hoarding converge with, or diverge from, digital curation within memory institutions. Compared to the DCC lifecycle, the practice of data hoarding begins with a focus on storage, rather than pre-ingest activities of appraisal and selection. Data-hoarders believe everything that you can get your hands on can and should be hoarded, and the only limit is how many TB hard drives you can buy. An interesting point of comparison is also seen in how hoarding addresses the DCC lifecycle's activities of *'create or receive'*; for data hoarding practice, these activities are much more active than passive, requiring considerable workarounds as data-hoarders seek out and capture materials through illicit means. Data-hoarders take a unique perspective on *'preservation actions'* compared to institutional practices concerned with migration, emulation, and fixity checks. Instead, data-hoarders preserve through torrents, seeding, and re-seeding, embracing the idea of, *'lots of copies keep stuff safe'*. From this perspective, access and preservation are much more closely aligned, or even seen as the same activity, enabled by their work outside of organizational policies and concerns with copyright.

We also observe that data hoarding practices are often explicitly driven by personal motivations of nostalgia and political stances. This presents an interesting contrast to curation practice in organizational settings, which can also, equally be influenced by personal interests of curators, though this may not be openly acknowledged due to a perceived need for institutional neutrality. The way data hoarding outwardly accepts the personal nature of memory work and collecting is a welcome change.

The key themes described above represent an initial analysis of how the r/DataHoarder community defines and constructs practices of data hoarding. We note that this preliminary work has several limitations, particularly in the selection of a dataset based on top posts. Many of the more everyday routines, as well as processes of enrollment and enculturation may not be present in our selected dataset. In future work we aim to study the *'long tail'* dataset of posts with significantly less engagement and employ ethnographic methods to generate rich descriptive observations to understand the values and ethics driving this practice. Additional posts studied being to reveal processes of identity formation through r/data-hoarder, seen in posts with sentiments such as *'this is why I hoard.'* This identity can also become all-consuming and challenging for posters who subsequently describe a need to stop hoarding or describe how hoarding affects the relationship with their significant other. The role of gender within this community, and the prevalence of pornography as a prevalent object of hoarding is another topic, we have not been able to analyze in detail in this preliminary work. Through subsequent phases of this study, we also hope to address if or how data hoarding distinguishes between data and other digital objects (like the articles, books, media, and games that are often the subject of hoarding). Ultimately, by comparing the work of data-hoarders with the work of curation in memory institutions, and what each is able to target in their collections, we aim to understand what gaps and absences remain.

## Conclusion

This work-in-progress highlights how r/DataHoarder offers a view of digital curation practice focused on the materiality of digital storage, and calls-to-action that raise public awareness of the potentially irrevocable loss of digital information. This presents useful counterpoints to commonly referenced and standardized practices of memory institutions by focusing on how communities conceptualize, practice, and develop their own ethical framework toward sustained access of digital materials. In future work we hope to address many questions and topics out-of-scope for this paper, such as expanding to other communities, comparing across communities, and applying both etic and emic coding to community posts.

## Acknowledgements

## About the authors

**Emily Maemura** is Assistant Professor at the School of Information Sciences, University of Illinois Urbana-Champaign, USA. She completed her PhD at the University of Toronto's Faculty of Information. Her research focuses on data practices and the activities of curation, description, characterization, and re-use of archived web data. She can be contacted at emaemura@illinois.edu

**Travis L. Wagner** is Assistant Professor at the School of Information Sciences, University of Illinois Urbana-Champaign, USA. They completed their PhD at the University of South Carolina's School of Information Science. Their research focuses on the sociotechnical relationships between archival artifacts, digital curation technologies, and community representation, with an emphasis on its impact on LGBTQIA+ communities. They can be contacted at wagnert@illinois.edu

## References

Basu, T. (2021, January 8). The scramble to archive Capitol insurrection footage before it disappears. MIT Technology Review. https://web.archive.org/web/20210108224223/https://www.technologyreview.com/2021/01/08/1015929/archive-capitol-insurrection-trump-maga-footage/

Bowker, G. C., & Star, S. L. (2000). Sorting Things Out: Classification and Its Consequences. MIT Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Britt, B. C., Britt, R. K., Hayes, J. L., & Oh, J. (2022). Continuing a community of practice beyond the death of its domain: Examining the Tales of Link subreddit. Behaviour & Information Technology, 41(1), 159–180. https://doi.org/10.1080/0144929X.2020.1797173

Brown, J. S., & Duguid, P. (1991). Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation. Organization Science, 2(1), 40–57.

Brown, J. S., & Duguid, P. (2001). Knowledge and Organization: A Social-Practice Perspective. Organization Science, 12(2), 198–213. https://doi.org/10.1287/orsc.12.2.198.10116

Chapman, E. M. (2023). ARCHIVING THE INSURRECTION: THE CASE OF R/DATAHOARDER. AoIR Selected Papers of Internet Research. https://doi.org/10.5210/spir.v2022i0.12987

Cox, A. (2005). What are communities of practice? A comparative review of four seminal works. Journal of Information Science, 31(6), 527–540. https://doi.org/10.1177/0165551505057016

Dallas, C. (2016). Digital curation beyond the "wild frontier": A pragmatic approach. Archival Science, 16(4), 421 - 457. https://doi.org/10.1007/s10502-015-9252-6

Grandhi, S. A., Plotnick, L., & Hiltz, S. R. (2020). An Internet-less World? Expected Impacts of a Complete Internet Outage with Implications for Preparedness and Design. Proc. ACM Hum.-Comput. Interact., 4(GROUP), 03:1-03:24. https://doi.org/10.1145/3375183

Graves, R. L., Perrone, J., Al-Garadi, M. A., Yang, Y.-C., Love, J., O'Connor, K., Gonzalez-Hernandez, G., & Sarker, A. (2022). Thematic Analysis of Reddit Content About Buprenorphine-

naloxone Using Manual Annotation and Natural Language Processing Techniques. Journal of Addiction Medicine, 16(4), 454–460. https://doi.org/10.1097/ADM.0000000000000940

Haythornthwaite, C., Kumar, P., Gruzd, A., Gilbert, S., Esteve del Valle, M., & Paulin, D. (2018). Learning in the wild: Coding for learning and practice on Reddit. Learning, Media, and Technology, 43(3), 219–235. https://doi.org/10.1080/17439884.2018.1498356

Hills, M. (2015). The expertise of digital fandom as a 'community of practice': Exploring the narrative universe of Doctor Who. Convergence, 21(3), 360–374. https://doi.org/10.1177/1354856515579844

Hudgins, W., Lynch, M., Schmal, A., Sikka, H., Swenson, M., & Joyner, D. A. (2020). Informal Learning Communities: The Other Massive Open Online 'C.' Proceedings of the Seventh ACM Conference on Learning @ Scale, 91–101. https://doi.org/10.1145/3386527.3405926

Kwon, K. H., Yu, W., Kilar, S., Shao, C., Broussard, K., & Lutes, T. (2020). Knowledge Sharing Network in a Community of Illicit Practice: A Cybermarket Subreddit Case. Proceedings of the 53rd Hawaii International Conference on Systems Sciences, 2731–2740. http://hdl.handle.net/10125/64076

Lave, J., & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511815355

LeBlanc, Z., Janco, A., Wermer-Colan, A., Dombrowski, Q., Kijas, A., Majstorovic, S., Strong, D., & Peaslee, E. (2022). A Conversation with the Organizers of Saving Ukrainian Cultural Heritage Online (SUCHO). Journal of Library Outreach and Engagement, 2(1), Article 1. https://doi.org/10.21900/j.jloe.v2i1.969

Lundeen, L. A., McCall, J. R., Bradshaw, A. S., LeBlanc, E. L., & Humphrey, L. (2024). Digital Catharsis or Harmful Exposure? A Thematic Analysis of Self-Directed Violence Reddit Posts. Social Media + Society, 10(3), 20563051241263562. https://doi.org/10.1177/20563051241263562

Maxwell, D., Robinson, S. R., Williams, J. R., & Keaton, C. (2020). "A Short Story of a Lonely Guy": A Qualitative Thematic Analysis of Involuntary Celibacy Using Reddit. Sexuality & Culture, 24(6), 1852–1874. https://doi.org/10.1007/s12119-020-09724-6

McMeekin, S., & Currie, A. (2022). DPC Competency Audit Toolkit Guide (1st ed.). Digital Preservation Coalition. https://doi.org/10.7207/dpccat22-01

Moran, J., Feltham, F., & Love, V. (2019). Building an Aotearoa New Zealand-wide Digital Curation Community of Practice. International Journal of Digital Curation, 14(1), Article 1. https://doi.org/10.2218/ijdc.v14i1.638

National Digital Stewardship Alliance (NDSA). (2019). 2019 Levels of Digital Preservation Matrix. https://osf.io/2mkwx/

Ogden, J. (2022). "Everything on the internet can be saved": Archive Team, Tumblr, and the cultural significance of web archiving. Internet Histories, 6(1–2), 1–20. https://doi.org/10.1080/24701475.2021.1985835

Panek, E. T. (2022). Understanding Reddit. Routledge.

Peltzman, S., Dietz, B., Butler, D., Walker, P., Farrell, J., Arroyo-Ramirez, E., Macquarie, C., Bolding, K., Helms, A., Venlet, J., Cobourn, A., Watson, C., Taylor, A., & Henke, K. (2022). Levels of Born-Digital Access. https://doi.org/10.17605/OSF.IO/R5F78

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. Social Media + Society, 7(2), 20563051211019004. https://doi.org/10.1177/20563051211019004

r/DataHoarder. (2017, July 30). A quick Datahoarder FAQ [Reddit Post]. https://web.archive.org/web/20171211044900/www.reddit.com/r/DataHoarder/comments/6qf716/a_quick_datahoarder_faq/

Rios, F., Lassere, M., Ruggill, J. E., & McAllister, K. S. (2020). Sustaining Software Preservation Efforts Through Use and Communities of Practice. International Journal of Digital Curation, 15(1), Article 1. https://doi.org/10.2218/ijdc.v15i1.696

Walker, D., Nost, E., Lemelin, A., Lave, R., & Dillon, L. (2018). Practicing environmental data justice: From DataRescue to Data Together. Geo: Geography and Environment, 5(2), e00061. https://doi.org/10.1002/geo2.61

Wenger, E. (1998). Communities of Practice: Learning as a Social System. Systems Thinker, 9(5), 2–3.

Wenger, E. (2000). Communities of Practice and Social Learning Systems. Organization, 7(2), 225–246. https://doi.org/10.1177/135050840072002