Unstable Data and the Unusual Case of the Prosody Excerpt in the Digital Library

Rebecca Sutton Koeser[1], Mary Naydan[1], and Meredith Martin[1,2]

[1] Center for Digital Humanities, Princeton University, Princeton, New Jersey, USA
[2] Department of English, Princeton University, Princeton, New Jersey, USA

Abstract

Stable data is essential for repeatable research, and cultural heritage data is an invaluable resource for computational humanities research, but the fluctuations within this kind of data pose challenges to reproducible, repeatable, and follow-up research. This paper uses the case study of HathiTrust Digital Library content within the Princeton Prosody Archive (PPA), particularly excerpted and augmented content, as a window into the surprising instability of this large-scale data. We analyze the rate of change in PPA excerpted content, all of PPA, and all of HathiTrust resulting from HathiTrust updates. We use this case study to illuminate the degree of change in HathiTrust, as one exemplar of a cultural heritage data aggregator, which we think is not well understood by computational researchers, and to consider the implications for research.

Keywords: humanities data, unstable data, reproducibility, digital libraries

1 Introduction

Data is essential for computational humanities research. Whether research is based on small-scale data to test a particular method, expansive collections for measuring larger trends across time, or something in between, stable data is crucial to interpreting results. The ability to share or recreate a dataset is equally crucial for reproducible research and to assess, critique, and take up new methodologies. However, the ability to build on previous research is difficult when there is instability in research data sources, particularly when that instability is unexpected or incompletely understood.

Stable data is particularly essential when it is used in novel or creative ways not anticipated by data publishers. Rufus Pollock, an early proponent of open knowledge and open data, has long argued that "the best thing to do with your data will be thought of by someone else" [26]. Computational humanists' reliance on cultural heritage data poses particular challenges, since GLAM (Galleries, Libraries, Archives, and Museums) sector workflows and goals are not always aligned with computational research. In this paper, we share insights gained from our discovery of surprisingly unstable data in the integrated and enhanced content from the HathiTrust Digital Library within the Princeton Prosody Archive (PPA). This case study provides an opportunity to reflect on the gaps in current standards for reproducible research, humanities and cultural heritage data publication, and computational humanities.[1]

In 2016, the FAIR Principles were published with the "ultimate goal ... to optimise the reuse of data" [7]. These guidelines include unique, persistent identifiers so data can be found, and documenting and structuring data according to community standards to support replication and combination. In 2017, the first statement on "Collections as Data" was published, with a goal

Rebecca Sutton Koeser, Mary Naydan, and Meredith Martin. "Unstable Data and the Unusual Case of the Prosody Excerpt in the Digital Library." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1345–1358. https://doi.org/10.63744/cTWwVSItf41f.

[1] Code and data associated with this paper is available at https://github.com/rlskoeser/chr2025-unstable-data-paper/

to "encourage computational use of digitized and born digital collections" [24]. Reusable data is an important step towards many of the modes of work that Christof Schöch elaborates within the conceptual space of "repetitive research" [28], but reusable data may not be sufficient. And data drawn from digital collections with greater "changeability and instability" require more detailed documentation on provenance and processing for reuse [3]. Many of the 2023 Collections as Data position statements offer glimpses of instability: items are accessioned, deaccessioned, and reappraised following curatorial workflows [15]; there is "considerable fluctuation" in adding and updating resources [22]; digital content should be treated as original, unique publications for preservation [27; 29], and yet these developments reflect the "destabilization of prior frameworks, theories, and practices" [1]. The FAIR principles do not mention data versioning, and provide no guidance on granularity of identifiers, but our experience indicates both are necessary.

2 Unstable data in the Princeton Prosody Archive

2.1 Princeton Prosody Archive

The Princeton Prosody Archive (PPA) is an open-source, full-text searchable database of 7,000+ English-language digitized works about the study of poetry, versification and pronunciation [19]. The works in the PPA — which include grammar books, elocution manuals, and scholarly articles, among many other kinds of books — were all published between 1532 and 1929, the current cut-off year for works in the public domain in the United States.

The PPA initially integrated only full volumes from HathiTrust, but was designed for expansion to other sources. It is one of the only digital humanities resources to present full-text OCR, page image thumbnails, and bibliographic metadata from multiple databases in a single dynamic interface.[2] The PPA also now supports excerpted works, such as book chapters or journal articles, to focus the collection only on relevant portions of larger works.

2.2 HathiTrust Digital Library

The majority of the works in the PPA come from HathiTrust Digital Library, a US-based, not-for-profit consortium of over 60 academic and research libraries from across North America and other countries. HathiTrust Digital Library began in 2008 with Google's mass digitization initiative and now provides "reading access" to its 18+ million digitized volumes "to the fullest extent allowable by U.S. and international copyright law" [11].

Similar large-scale aggregators, often led by national libraries, exist in Europe and beyond. **TROVE**, maintained by the National Library of Australia, collects billions of digital items from Australian libraries, universities, museums, galleries, and archives.[3] **Gallica**, the digital library of the Bibliothèque nationale de France (BnF) and over 300 partners, provides access to over 10 million items.[4] **Europeana** makes available various kinds of media and accompanying metadata from thousands of cultural institutions.[5] Smaller cultural heritage aggregators include Switzerland's **e-rara**[6] and **e-manuscripta**[7]and Finland's **Finna**.[8] We mention these aggregators to suggest that the kinds of data instability we encountered in HathiTrust will be found in *any* large-scale cultural heritage aggregator because of GLAM workflows to manage and improve collections, as well as

[2] The PPA now also includes content from Gale/Cengage's Eighteenth Century Collections Online (ECCO) and Early English Books Online using editions from the Text Creation Partnership (EEBO-TCP).
[3] `https://trove.nla.gov.au`
[4] `https://gallica.bnf.fr`
[5] `https://www.europeana.eu`
[6] `https://www.e-rara.ch`
[7] `https://www.e-manuscripta.ch`
[8] `https://finna.fi`

the ephemeral nature of technical infrastructure. These changes often involve a trade-off between reproducibility and improvement. For example, TROVE's crowdsourced text-correcting feature improves OCR transcripts, but means that researchers will be working with different transcripts depending on the date of download [20]. This opportunity for continued content enhancement is valuable, but depending on the extent of the changes, could have downstream effects on word-level analysis. As another familiar example, Europeana's metadata coordinators explain to their users why they might find broken links and how some of these can never be fixed [33].

HathiTrust has a long history of enthusiastic collaboration with researchers and computational humanists. Many projects focused on methods for, first, finding certain material within HathiTrust's massive digital library, and then testing text and data mining methods on that collection using the tools and services provided by HathiTrust Research Center (HTRC) [9]. Prominent examples of this two-step process include Gioia Stevens's **Early American Cookbooks** project (now a permanent collection within HathiTrust) [30] and Laure Thompson and David Mimno's **20th Century English-Language Speculative Fiction** (a recommended workset within HTRC) [12; 31]. Recent work from Ryan Dubnicek and collaborators has focused on using machine learning methods to discover particular genres within HathiTrust Digital Library [4; 25] and building out the customizable capacities of HTRC's Extracted Features dataset through the development of an open API [34]. The content is limited to HathiTrust items, and the algorithms, datasets, and advanced computing environments are all contained within HTRC's platform.[9]

2.3 PPA's unusual use of HathiTrust

Unlike most other projects, the PPA pulls content dynamically from multiple HathiTrust systems and combines it with content from non-HathiTrust sources, cutting across the boundaries of digital resources and adding a layer of technical complexity.[10] PPA uses HathiTrust's API to import bibliographic metadata, which it presents alongside page-level data (OCR) from local copies of METS XML and plain text files accessed via rsync and page image thumbnails dynamically loaded from HathiTrust's image server.

In addition, the PPA enhances the data pulled from HathiTrust, particularly for excerpts. Whereas HathiTrust indexes full works and journal volumes, the PPA enables searching and browsing at the more granular level of the book chapter or journal article. The bibliographic metadata for prosody-related excerpts had to be identified and curated manually [21], providing a level of specificity not available in HathiTrust.

While HathiTrust was an enthusiastic partner and participated in several rounds of negotiations about how PPA would access their data, it eventually became clear — especially through the case of the prosody excerpt — that their infrastructure was not designed to be used by external projects in this persistent, dynamic manner.

2.3.1 The problem of excerpts

The instability of HathiTrust's data was first evidenced when excerpt-level functionality was added to the PPA. Suddenly, there were mismatches between the thumbnail images and full-text snippets.

[9] In October 2024, HathiTrust announced that it will be shutting down HTRC because "many of our members have rarely utilized HTRC's services," leaving the future of TDM on HathiTrust works uncertain [6].

[10] It is difficult to identify other digital projects with unusual uses of data or technical infrastructure; this requires insider knowledge of the processes and sources used to build projects, which is less often narrated than research methods, analytical results, or final outputs. Meredith Martin's new monograph, *Poetry's Data*, is a notable exception to this trend and argues for the value of narrating the process of navigating our digital research infrastructures [18]. The Princeton Geniza Project and the Shakespeare & Company Project are two additional examples of unexpected use: both projects attach research data to digital images published by GLAM institutions via International Image Interoperability Framework (IIIF) APIs [16; 19]. This kind of persistent use is within the guidelines of the published IIIF APIs, but atypical.

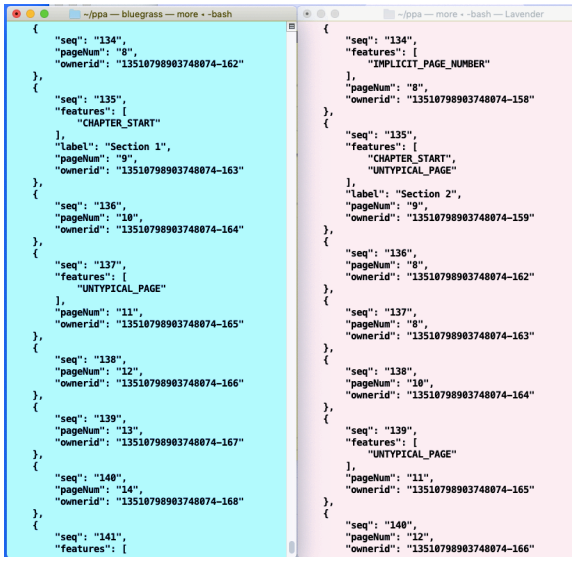


Figure 1: Side-by-side screenshots showing page-level metadata for "The Scansion of Middle English Alliterative Verse" by William E. Leonard [17], generated from the page's HTML and JavaScript, on two different dates: March 18, 2024 (left) and March 26, 2024 (right). Note the change in `pageNum` beginning at `seq` 136.

Page ranges were incorrect. How was this possible? We realized that because the PPA was regularly pulling content from multiple HathiTrust systems at different intervals, HathiTrust's routine rescanning of its material could alter the digital page numeration that was used to specify excerpt ranges.

We suspect that HathiTrust's regular rescanning is not widely known or even fully understood among researchers. The main public indication is the following disclaimer on an "About" page (emphasis added): "*The HathiTrust collection is not static.* Works get added to the collection every day, and *sometimes a digital item may be updated with a new version*. Bibliographic records can be updated when contributors send us corrections. Copyright and access statuses may change as items undergo copyright review or we receive permissions agreements from copyright holders" [10].

To address this problem, we sought a solution to automate fixing range changes whenever HathiTrust made updates. At first glance, this task seemed computationally tractable. Figure 1 shows the page-level metadata for the same set of pages on two different dates and the types of changes that occur. Some of the changes involve recharacterizing page-level semantic metadata (e.g., flagging `UNTYPICAL_PAGE` or renumbering section starts), though it is worth noting that the updated attributes are not always more accurate. For PPA excerpts, we were most interested in the relative changes between the `seq` field (digital page index) and the `pageNum` field (page number in the source edition).

Digging deeper, we discovered complex edge cases in page numeration that would make automation difficult, if not impossible, such as articles with *three* different page numbers for a single page (see Figure 2).

After the technical cost of automatically fixing page ranges became apparent, the PPA team resolved that manual correction was the best solution.

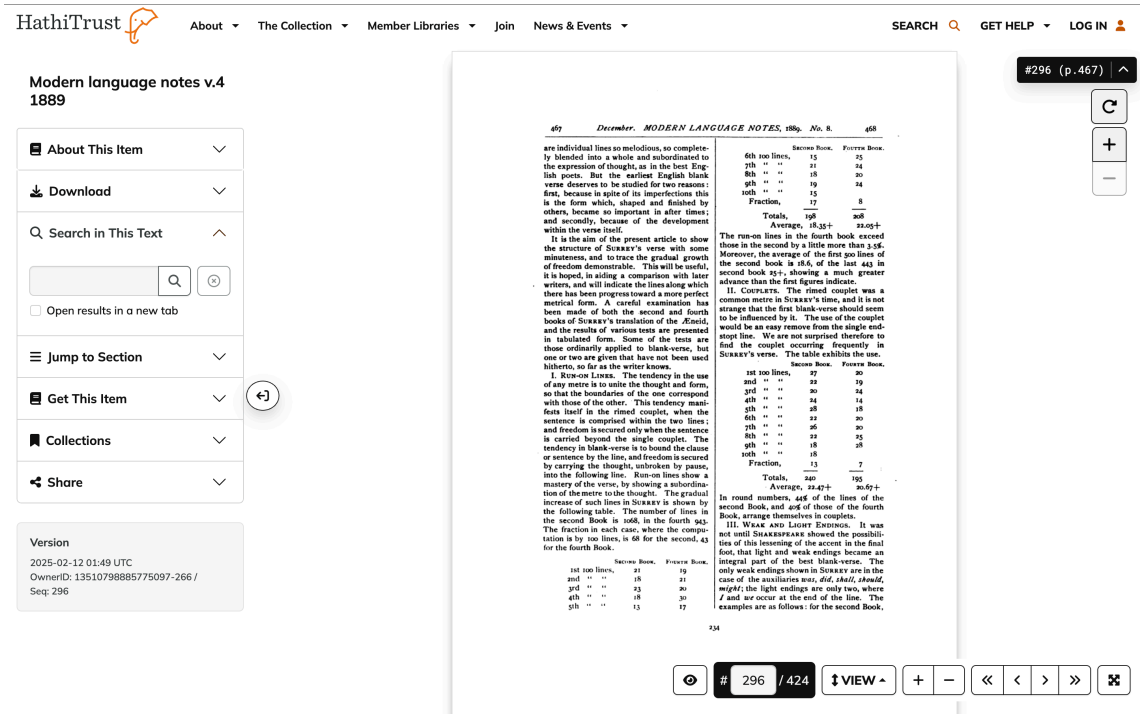


Figure 2: The above page from O. F. Emerson's "The Development of Blank Verse: A Study of Surrey" in the journal *Modern Language Notes* [5] displays three different page numbers. Each column of *MLN* gets its own page number (left, 467; right, 468), and each page gets a third, found at the bottom center (234). HathiTrust's black navigation controls show that it takes the upper left page number as the original/physical page number (p. 467). This convention is not the most intuitive, and it means that scrolling through HathiTrust's reader, the original page numbers increase by two instead of one. The digital page number/sequence number (#296) is based on scanning and subject to change with rescanning.

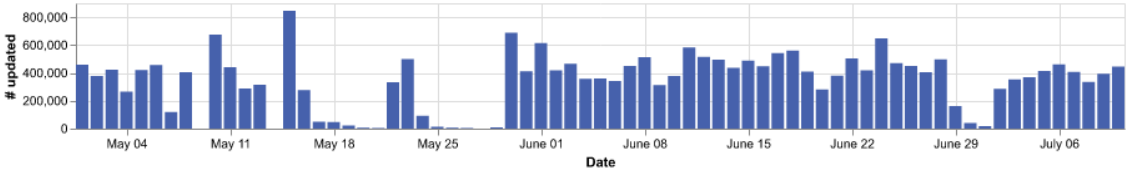


Figure 3: Number of volumes updated daily in all of HathiTrust from May 1 to July 10, 2025.

3 Quantifying HathiTrust's rate of change

3.1 Changes in PPA excerpts

A November 2023 spreadsheet used by the PPA team to correct excerpt page ranges provides a window into the degree of instability. After we discovered that the PPA was pulling erroneous content for some excerpts, we tasked a student researcher with manually checking and correcting the page ranges. We discovered that out of 517 total excerpts, 121 (23.4%) had changed over the course of two years. For 10 of those, the length of the excerpt changed, in most cases by two pages.[11] This might occur if, for instance, a digital edition included a duplicated page, which was later found and removed. Excerpt starting pages shifted an average of 6.4 pages; 9 volumes shifted by more than 10 pages, and one extreme outlier, Thomas Stewart Omond's "Swinburne as a Metrician" in volume 76 (1909) of *The Academy and literature* [23], shifted by 240 pages! This could happen if HathiTrust replaced a partial scan with a more complete one. Most troubling, when we analyze the overlap between pages in range before and after these updates, there are 36 excerpts with *no pages in common* between the initial and updated page range; that is, without updating the page range, we would be including *none* of the intended contents. While small shifts in page numeration are negligible for a massive collection like HathiTrust, for smaller, curated, scholar-focused projects, even small changes like these can have big implications, leading to errors in scholarship, confusion among users, and distrust in the quality of the resource.

3.2 Changes in all of HathiTrust

Observing this volatility at the small scale of PPA excerpts led us to wonder about the rate of change for PPA content more generally and in HathiTrust at scale. Fortunately for us, HathiTrust is incredibly transparent about changes to its collection — at least at a high level. In addition to emails HathiTrust regularly sends out with lists of records that are no longer public domain, which dataset users must remove from their copy of HathiTrust data (see Appendix A and B), HathiTrust publishes monthly and daily files with data about updated records. They note that a volume might be included in this update file for *any* of the following reasons: it was newly deposited, an existing item was replaced with a new copy, there were changes to rights and access permissions, or the bibliographic metadata was updated [8]. While the occurrence of an update is therefore relatively easy to track, identifying the exact *type* of change is not.[12] Figure 3 charts the number of updated volumes across all of HathiTrust from May 1 to July 10, 2025. The number of updates varies

[11] There were two outliers with differences in length of 25 and 16 pages; preliminary investigations suggest that these were due to data errors in the original excerpts. For instance, one excerpt was originally only pulling a single page when the article was actually several pages long; most likely, someone neglected to input the end range. However, it is difficult to confirm with certainty after the fact.

[12] Our investigations made it clear that the modification time of the METS XML file corresponds to the date in the public version label on the full view page for individual items; the text content could be modified before or after the METS XML. We also investigated the preservation metadata (PREMIS) included in the METS file for each volume, looking at volumes for specific PPA excerpts which did and did not shift (§3.1). This log documents data maintenance and data migration events, and while they may indicate that a volume has changed, they do not provide additional clarity on the specifics of that change.

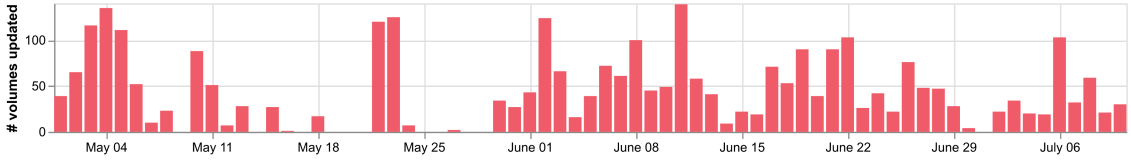


Figure 4: Number of PPA HathiTrust volumes updated daily from May 1 to July 10, 2025.

on any given day from nearly negligible to more than 800,000 volumes. The day with the most updates during this time period was May 15, 2025 with 846,329 volumes updated (4.5% of all of HathiTrust). The average daily update over this period is 348,392, which is only 1.8% of all of HathiTrust.

3.3 Changes in all of PPA

What about all the HathiTrust volumes within the Princeton Prosody Archive? How frequently are they changing? One way to answer this question is to filter the data on HathiTrust updates to volumes included in PPA; when we do, we find that PPA updates track somewhat with the larger updates (refer to Figure 4 and compare with Figure 3). Although these changes are on a much smaller scale, there are still multiple days when more than a hundred volumes changed.

Another way to investigate this question is by comparing two different snapshots of the PPA full-text corpus. We created one version that was last updated November 2024 and another in February 2025, a difference of roughly three months.[13] There are slight differences between the number of pages in these two versions, with a little over 1.5 million pages in common. When we match pages based on exactly equal text contents within the same volume (excluding pages with no text), we find 855,930 (55.8%) matching pages across the two corpora; of those, 14,423 pages (1.7% of the total 1.5 million pages) have shifted sequence within their volume, indicating a change in structure with no update to the text. The mismatches for the remaining pages (44.2% of our corpus) are likely due to OCR changes. While it is appealing to think that the content has been improved, this means that any word or token-level analysis on the prior version of the corpus would need to be either rerun or realigned to the updated text.[14]

3.4 Implications

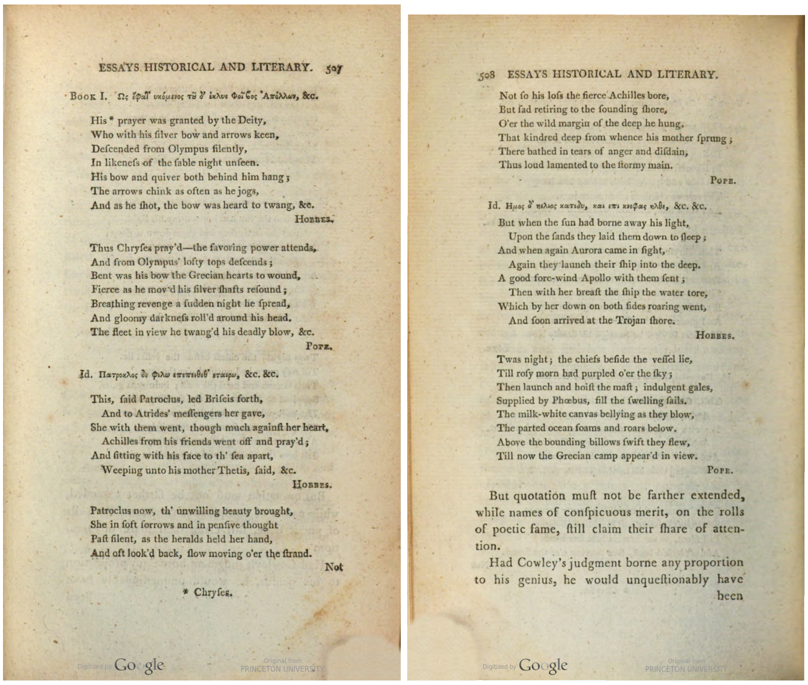
The difficulty of dealing with HathiTrust's frequent changes factored into the PPA team's decision to shift from maintaining a dynamic database to extracting the data and making the full-text corpus useful for computational analysis instead. One project aimed to detect and identify lines of poetry quoted across the million+ pages in the PPA. Systematically identifying these poems was a first step towards answering research questions about English prosody; a dataset of poem excerpts could illuminate when and how particular poets or poems became the exemplars for poetic forms or figures of speech, or how quickly after publication a poet's work becomes a canonical example, or the network of examples being reused from other prosodists.

As we considered approaches to poetry detection within PPA, one collaborator suggested leveraging Ted Underwood's dataset of page-level genre predictions for HathiTrust volumes [32]. This dataset includes such granular predictions because, as the researchers note, volumes are rarely a

[13] The earlier version is the corpus that has been used as the basis for computational research on PPA; the later one was generated as part of a matching text and image snapshot that was provided as a one-time dataset by HathiTrust.

[14] The PPA's ECCO data provides an interesting counterpoint to the instability of HathiTrust: the data itself is unchanged, and the OCR text for these materials have not been updated since 2008, in spite of the known poor quality of the text and substantial improvements in OCR technology since the collection was originally digitized [13].

single genre and often collect disparate materials. While the genre prediction task is not equivalent to poetry detection, there is some overlap in goals. Pages labeled as poetry could be used as a starting point or confirmation of results from other methods, an approach which would fall under reuse of data or follow-up research, in Schöch's terminology [28]. However, the degree of change possible in HathiTrust materials means that this page-level data is basically unusable. As a demonstration of this problem, we offer one example: the essay "On Stile and Versification" by William Belsham [2] appears in Underwood's dataset and in the PPA. Figure 5 shows the only two pages in the essay that are majority poetry and discusses the discrepancies in page sequence numbers between the two datasets. Figure 6 shows the very brief poetry excerpts that occur in the rest of the essay and discusses how they are characterized.


ESSAYS HISTORICAL AND LITERARY. 507

Book I. Ὡς ἔφατ' εὐχόμενος τοῦ δ' ἔκλυε Φοῖβος Ἀπόλλων, &c.

His * prayer was granted by the Deity,
Who with his ſilver bow and arrows keen,
Deſcended from Olympus ſilently,
In likeneſs of the ſable night unſeen.
His bow and quiver both behind him hang;
The arrows chink as often as he jogs,
And as he ſhot, the bow was heard to twang, &c.
HOBBES.

Thus Chryſes pray'd—the favoring power attends,
And from Olympus' lofty tops deſcends;
Bent was his bow the Grecian hearts to wound,
Fierce as he mov'd his ſilver ſhafts reſound;
Breathing revenge a ſudden night he ſpread,
And gloomy darkneſs roll'd around his head.
The fleet in view he twang'd his deadly blow, &c.
POPE.

Id. Πατροκλος δε φιλω επεπειθεθ' εταιρω, &c. &c.

This, ſaid Patroclus, led Briſeis forth,
And to Atrides' meſſengers her gave,
She with them went, though much againſt her heart,
Achilles from his friends went off and pray'd;
And ſitting with his face to th' ſea apart,
Weeping unto his mother Thetis, ſaid, &c.
HOBBES.

Patroclus now, th' unwilling beauty brought,
She in ſoft ſorrows and in penſive thought
Paſt ſilent, as the heralds held her hand,
And oft look'd back, ſlow moving o'er the ſtrand.
Not

* Chryſes.

Digitized by Google — Original from PRINCETON UNIVERSITY

508 ESSAYS HISTORICAL AND LITERARY.

Not ſo his loſs the fierce Achilles bore,
But ſad retiring to the ſounding ſhore,
O'er the wild margin of the deep he hung,
That kindred deep from whence his mother ſprung;
There bathed in tears of anger and diſdain,
Thus loud lamented to the ſtormy main.
POPE.

Id. Ημος δ' ηελιος κατεδυ, και επι κνεφας ηλθε, &c. &c.

But when the ſun had borne away his light,
Upon the ſands they laid them down to ſleep;
And when again Aurora came in ſight,
Again they launch their ſhip into the deep.
A good fore-wind Apollo with them ſent;
Then with her breaſt the ſhip the water tore,
Which by her down on both ſides roaring went,
And ſoon arrived at the Trojan ſhore.
HOBBES.

Twas night; the chiefs beſide the veſſel lie,
Till roſy morn had purpled o'er the ſky;
Then launch and hoiſt the maſt; indulgent gales,
Supplied by Phœbus, fill the ſwelling ſails.
The milk-white canvas bellying as they blow,
The parted ocean foams and roars below.
Above the bounding billows ſwift they flew,
Till now the Grecian camp appear'd in view.
POPE.

But quotation muſt not be farther extended, while names of conſpicuous merit, on the rolls of poetic fame, ſtill claim their ſhare of attention.

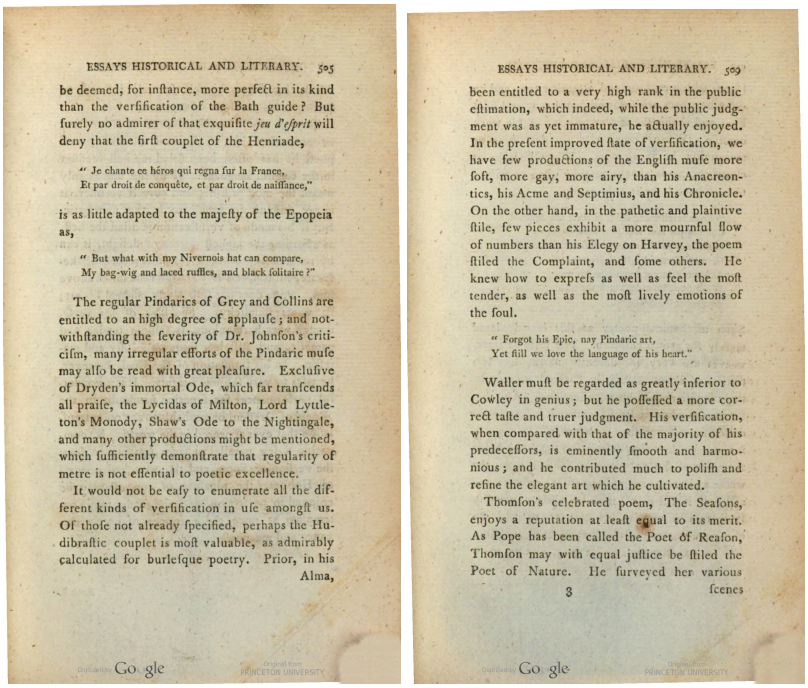
Had Cowley's judgment borne any proportion to his genius, he would unqueſtionably have been

Digitized by Google — Original from PRINCETON UNIVERSITY

Figure 5: The only two pages in this essay that are majority poetry. Underwood's dataset could potentially help us find poem excerpts like these. However, in the PPA's updated dataset, the digital sequence numbers for these two pages are 515 and 516. In Underwood's page-level genre metadata, the only two pages labeled "poetry" are digital pages 510 and 511. Images courtesy of HathiTrust.

This example shows that page-level genre predictions have potential for research questions within PPA or other projects based on HathiTrust materials, but updates in HathiTrust content make them unusable for work on different versions of the data.

Like Underwood, the PPA project team is using digital sequence to refer to pages in the found-poems dataset. What do we do with the fact that this dataset, like Underwood's, refers to a snapshot of HathiTrust and the PPA at a moment in time? For researchers working with HathiTrust data, there is no stable page identifier across time. A stable page ID for HathiTrust is not feasible due to the scale of labor involved in creating and maintaining HathiTrust, since there are so many individual libraries with their own workflows feeding into the aggregator. While this lack of stable referents is therefore understandable, it nevertheless poses a problem for computational research and reproducibility that the field has yet to solve.



ESSAYS HISTORICAL AND LITERARY. 505

be deemed, for inſtance, more perfect in its kind than the verſification of the Bath guide? But ſurely no admirer of that exquiſite *jeu d'eſprit* will deny that the firſt couplet of the Henriade,

" Je chante ce héros qui regna ſur la France,
Et par droit de conquête, et par droit de naiſſance,"

is as little adapted to the majeſty of the Epopeia as,

" But what with my Nivernois hat can compare,
My bag-wig and laced ruffles, and black ſolitaire?"

The regular Pindarics of Grey and Collins are entitled to an high degree of applauſe; and notwithſtanding the ſeverity of Dr. Johnſon's criticiſm, many irregular efforts of the Pindaric muſe may alſo be read with great pleaſure. Excluſive of Dryden's immortal Ode, which far tranſcends all praiſe, the Lycidas of Milton, Lord Lyttleton's Monody, Shaw's Ode to the Nightingale, and many other productions might be mentioned, which ſufficiently demonſtrate that regularity of metre is not eſſential to poetic excellence.

It would not be eaſy to enumerate all the different kinds of verſification in uſe amongſt us. Of thoſe not already ſpecified, perhaps the Hudibraſtic couplet is moſt valuable, as admirably calculated for burleſque poetry. Prior, in his Alma,

ESSAYS HISTORICAL AND LITERARY. 509

been entitled to a very high rank in the public eſtimation, which indeed, while the public judgment was as yet immature, he actually enjoyed. In the preſent improved ſtate of verſification, we have few productions of the Engliſh muſe more ſoft, more gay, more airy, than his Anacreontics, his Acme and Septimius, and his Chronicle. On the other hand, in the pathetic and plaintive ſtile, few pieces exhibit a more mournful flow of numbers than his Elegy on Harvey, the poem ſtiled the Complaint, and ſome others. He knew how to expreſs as well as feel the moſt tender, as well as the moſt lively emotions of the ſoul.

" Forgot his Epic, nay Pindaric art,
Yet ſtill we love the language of his heart."

Waller muſt be regarded as greatly inferior to Cowley in genius; but he poſſeſſed a more correct taſte and truer judgment. His verſification, when compared with that of the majority of his predeceſſors, is eminently ſmooth and harmonious; and he contributed much to poliſh and refine the elegant art which he cultivated.

Thomſon's celebrated poem, The Seaſons, enjoys a reputation at leaſt equal to its merit. As Pope has been called the Poet of Reaſon, Thomſon may with equal juſtice be ſtiled the Poet of Nature. He ſurveyed her various

3 ſcenes

Figure 6: Two pages with short poetry excerpts. In Underwood's dataset, pages like these are classified as "nonfiction prose" rather than poetry. Underwood's dataset would therefore not help us find short poetry excerpts like these. Images courtesy of HathiTrust.

4 Conclusions

There are many different methods and approaches for computational research on large-scale corpora and digital collections. Because of the complexity of reconciling augmented, curated PPA data with ongoing HathiTrust updates, we have followed the familiar pattern of working with a frozen snapshot of a corpus to test our methods and generate a new derivative dataset, which will be the basis for additional research. It is important to have research results linked to a specific version of a dataset, which can be cited and potentially shared, depending on permissions. However, this approach has downsides; even within our own project, our desires to refine the dataset and enhance the metadata run the risk of introducing unintentional changes into the corpus and becoming out of sync with work derived from the previous version. These challenges are even more substantial beyond the bounds of an individual project, making it difficult to share research results for reuse, critical review, follow-up research, or to feed back into enhancing and augmenting critical heritage datasets and digital collections.

Ideally, cultural heritage data and computational research constitute a virtuous cycle: data is used to test new methods or algorithms and to make domain-specific arguments, and results feed back into improvements to data, methods, and new scholarly arguments through repeated and follow-up research. GLAM institutions and researchers are often related but not perfectly aligned in their goals, and researchers may often require different scales, nuance, and complexity than can be supported by curatorial workflows. A crucial first step is to be transparent about the degree and scope of instability in data, and to publish and use versioned data whenever feasible. FAIR does not go far enough; perhaps expanding to include ethics and expertise, more detailed source information, and time stamps ('FAIREST'), as proposed by Béatrice Joyeux-Prunel [14], would go farther. We must develop new community standards, whether for versioning or more granular identifiers, as well as tools or methods to work more robustly with unstable data for reproducible, repeatable research.

Acknowledgements

This work is supported by the Center for Digital Humanities at Princeton University. Thanks to Natalia Ermolaev for reviewing and providing feedback that resulted in a stronger argument and clearer prose. Thanks to Brian Kernighan for taking such a keen interest in this unstable data problem and sharing with us the process and results of his exploratory investigations (or "sleazy hacks", as he called them!), which include the screenshots in Figure 1. Thanks to our former and current collaborators at HathiTrust, especially Sandra McIntyre, Eleanor Koehl, and Ryan Dubnicek, for their support of the Princeton Prosody Archive.

References

[1] Bailey, Jefferson. "Collections as Data as Destabilization". In: *Collections as Data: Part to Whole Final Report*, ed. by Padilla Thomas, Hannah Scates Kettler, and Yasmeen Shorish. Zenodo, Nov. 2023. DOI: https://doi.org/10.5281/zenodo.10161976.

[2] Belsham, William. "On Stile and Versification". In: *Essays philosophical and moral, historical and literary*. Vol. v.2. London: G.G. and J. Robinson, 1799, pp. 482–513. URL: https://prosody.princeton.edu/archive/njp.32101076530979-p482/.

[3] Burrows, Toby. "Reproducibility, verifiability, and computational historical research". In: *International Journal of Digital Humanities 5*, no. 2 (2023), pp. 283–298. DOI: 10.1007/s42803-023-00068-9.

[4] Dubnicek, Ryan and Underwood, Ted. "Piloting A Machine Learning Approach to Identify English-Language Fiction in the HathiTrust Digital Library". Paper presented at Digital Humanities 2023 conference. Graz, Austria, July 2023.

[5] Emerson, O. F. "The Development of Blank Verse: A Study of Surrey". In: *Modern language notes 4* (1889), pp. 466–472. URL: https://prosody.princeton.edu/archive/mdp.39015060429746-p466/.

[6] Furlough, Mike. "Plans for the HathiTrust Research Center". Oct. 2024. URL: https://www.hathitrust.org/press-post/plans-for-hathitrust-research-center/.

[7] GO FAIR. "FAIR Principles". Online. 2019. URL: https://www.go-fair.org/fair-principles/.

[8] HathiTrust. "Hathifiles". Online. 2023. URL: https://www.hathitrust.org/member-libraries/resources-for-librarians/data-resources/hathifiles/.

[9] HathiTrust. "HathiTrust Research Center". Online. 2023. URL: https://www.hathitrust.org/about/research-center/.

[10] HathiTrust. "How to Use HathiTrust Data Resources". Online. 2023. URL: https://www.hathitrust.org/member-libraries/resources-for-librarians/data-resources/.

[11] HathiTrust. "Welcome to HathiTrust". Online. 2023. URL: https://www.hathitrust.org/about/.

[12] HathiTrust Research Center. "Recommended worksets". Online. 2021. URL: https://analytics.hathitrust.org/staticrecommendedworksets.

[13] Hill, Mark J and Hengchen, Simon. "Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study". In: *Digital Scholarship in the Humanities 34*, no. 4 (Dec. 2019), pp. 825–843. DOI: 10.1093/llc/fqz024.

[14] Joyeux-Prunel, Béatrice. “Digital humanities in the era of digital reproducibility: towards a fairest and post-computational framework”. In: *International Journal of Digital Humanities 6*, no. 1 (Apr. 2024), pp. 23–43. DOI: 10.1007/s42803-023-00079-6.

[15] Knazook, Beth and Narlock, Mikala. “Building the collections of tomorrow”. In: *Collections as Data: Part to Whole Final Report*, ed. by Padilla Thomas, Hannah Scates Kettler, and Yasmeen Shorish. Zenodo, Nov. 2023. DOI: https://doi.org/10.5281/zenodo.10161976.

[16] Kotin, Joshua and Koeser, Rebecca Sutton. “Shakespeare and Company Project”. Publisher: Center for Digital Humanities, Princeton University. 2021. URL: https://shakespeareandco.princeton.edu/.

[17] Leonard, William E. “The Scansion of Middle English Alliterative Verse”. In: *Studies in language and literature v.8-11* (1920), pp. 58–104. URL: https://prosody.princeton.edu/archive/uiug.30112046384886-p58/.

[18] Martin, Meredith. *Poetry's data : digital humanities and the history of prosody*. Princeton: Princeton University Press, 2025.

[19] Martin, Meredith, Wilson, Meagan, Naydan, Mary, and Koeser, Rebecca Sutton. “Princeton Prosody Archive”. Publisher: Center for Digital Humanities, Princeton University. 2018. URL: https://prosody.princeton.edu/.

[20] National Library of Australia. “Text Correction”. Online. 2025. URL: https://trove.nla.gov.au/help/your-trove-tools/text-correction.

[21] Naydan, Mary and Koeser, Rebecca Sutton. “Book Excerpts, Journal Articles, and Better Metadata”. In: *Princeton Prosody Archive* (Aug. 2024), ed. by Meredith Martin. Publisher: Center for Digital Humanities, Princeton University. URL: https://prosody.princeton.edu/editorial/2024/08/book-excerpts-journal-articles-and-better-metadata/.

[22] Neudecker, Clemens. “Collections as data for machine learning: if we build it, who will come”. In: *Collections as Data: Part to Whole Final Report*, ed. by Padilla Thomas, Hannah Scates Kettler, and Yasmeen Shorish. Zenodo, Nov. 2023. DOI: https://doi.org/10.5281/zenodo.10161976.

[23] Omond, Thomas Stewart. “Swinburne as a Metrician”. In: *The Academy and literature 76* (1909), pp. 32–33. URL: https://prosody.princeton.edu/archive/uc1.c2641998-p32/.

[24] Padilla, Thomas, Allen, Laurie, Varner, Stewart, Potvin, Sarah, Roke, Elizabeth Russey, and Frost, Hannah. “The Santa Barbara Statement on Collections as Data”. Online. 2018. URL: https://collectionsasdata.github.io/statement/.

[25] Parulian, Nikolaus Nova, Dubnicek, Ryan, Worthey, Glen, Evans, Daniel J., Walsh, John A., and Downie, J. Stephen. “Uncovering Black Fantastic: Piloting A Word Feature Analysis and Machine Learning Approach for Genre Classification”. In: *Proceedings of the Association for Information Science and Technology*. Vol. 59. 1. 2022, pp. 242–250. DOI: 10.1002/pra2.620.

[26] Pollock, Rufus. “Open Data: a means to an end, not an end in itself”. Online. 2011. URL: https://blog.okfn.org/2011/09/15/open-data-a-means-to-an-end-not-an-end-in-itself/.

[27] Roke, Elizabeth Russey. “Moving Collections as Data from a Patchwork of Projects to a Sustainable Program”. In: *Collections as Data: Part to Whole Final Report*, ed. by Padilla Thomas, Hannah Scates Kettler, and Yasmeen Shorish. Zenodo, 2023. DOI: https://doi.org/10.5281/zenodo.10161976.

[28] Schöch, Christof. “Repetitive research: a conceptual space and terminology of replication, reproduction, revision, reanalysis, reinvestigation and reuse in digital humanities”. In: *International Journal of Digital Humanities 5*, no. 2 (Nov. 2023), pp. 373–403. DOI: 10.1007/s42803-023-00073-y.

[29] Steven, Claeyssens. “No Size Fits all: Publishing Collections as Data”. In: *Collections as Data: Part to Whole Final Report*, ed. by Padilla Thomas, Hannah Scates Kettler, and Yasmeen Shorish. Zenodo, 2023. DOI: https://doi.org/10.5281/zenodo.10161976.

[30] Stevens, Gioia. “New Metadata Recipes for Old Cookbooks: Creating and Analyzing a Digital Collection Using the HathiTrust Research Center Portal”. In: *The Code4Lib Journal* , no. 37 (July 2017). URL: https://journal.code4lib.org/articles/12548.

[31] Thompson, Laure and Mimno, David. “Building Large-Scale Collections of Genre Fiction: Final Report”. Online. 2020. URL: https://laurejt.github.io/papers/htrc-acs-final-report.pdf.

[32] Underwood, Ted. “Page-Level Genre Metadata for English-Language Volumes in HathiTrust, 1700-1922”. Figshare. 2014. DOI: 10.6084/m9.figshare.1279201.

[33] Vos, Marjolein de. “Keeping digitised heritage accessible: the case of broken links”. Online. 2018. URL: https://pro.europeana.eu/post/keeping-digitised-heritage-accessible-the-case-of-broken-links.

[34] Walsh, John A., Layne-Worthey, Glen, Jett, Jacob, Capitanu, Boris, Organisciak, Peter, Dubnicek, Ryan, and Downie, J. Stephen. ““The library is open!”: Open data and an open API for the HathiTrust Digital Library”. In: *Proceedings of the Computational Humanities Research Conference 2023*, ed. by Artjoms Šeļa, Fotis Jannidis, and Iza Romanowska. Vol. 3558. CEUR Workshop Proceedings. Dec. 2023, pp. 703–714.

A HathiTrust Statement for Dataset Distribution

By my signature, I acknowledge and confirm the following:

1. I am receiving texts from the University of Michigan that are made available under an agreement between my sponsoring institution - [indicate sponsoring institution, e.g., Dartmouth College] - and Google.
2. I have read this agreement and agree to abide by its terms and to use the texts in accordance with the statement of my research, as submitted to the University of Michigan.
3. I agree to notify the University of Michigan of any changes that are made in the scope or nature of my research.
4. I understand that volumes I receive from the University of Michigan may be determined at a later date to be in copyright. I agree to delete these volumes and any copies that have been made upon notification from the University of Michigan. I agree to notify the University of Michigan at feedback@issues.hathitrust.org to confirm deletion of any such volumes.

Name	Signature	Date
Title		
Email	Phone	

B Example HathiTrust deletion email for public domain dataset

```
Subject: Delete notifications for ht_text_pd dataset
From: HathiTrust <support@hathitrust.org>

Dear HathiTrust dataset recipient,

This email is to notify you that volumes in the HathiTrust
"ht_text_pd" dataset, of which you have downloaded all
or a subset of files, no longer meet the criteria for
inclusion in the dataset, and you no longer are allowed to
use them in your research.

Please review the data you have synced from HathiTrust to
check whether you have the volumes listed below. If so,
delete all copies you retain of these volumes in
accordance with our terms of use. Alternatively, you may
delete your copy of the dataset and re-sync to the updated
dataset.

If you no longer possess HathiTrust datasets, or if you
have other questions regarding datasets, then please email
support@hathitrust.org.

Thank you,

HathiTrust

===BEGIN ID LIST===
[ids omitted]
...
...
...
===END ID LIST===
```