Journal of Open Humanities Data

The Open Pynchon Reception Corpus: From Bibliographic Metadata to Reusable Scholarly Full Text

DATA PAPER

ERIK KETZAN

ubiquity press

ABSTRACT

This paper introduces the *Open Pynchon Reception Corpus* (OPRC), a collection of 52 scholarly literary studies (597,608 word tokens) on the texts of American author Thomas Pynchon (1937–), which are licensed under Creative Commons Attribution (CC BY) and published between 2012 and 2025. It extends a prior project, the *Thomas Pynchon Online Bibliography* (TPOB), an open resource containing metadata for over 2,500 Pynchon studies. OPRC is a corpus based on the subset of these scholarly studies that are available under CC BY licences. The corpus release contains the original PDFs (and, for one text, the saved publisher HTML), the derived .txt files, and bibliographic metadata, together with a licence-evidence audit, extraction and validation scripts, and SHA-256 checksum manifests. OPRC supports computational literary studies of Pynchon's reception, including quotation and text-reuse studies, and its design and CC BY criterion support transparent and replicable reuse.

CORRESPONDING AUTHOR:

Erik Ketzan

Department of Digital Humanities, King's College London, London, United Kingdom

erik.ketzan@kcl.ac.uk

KEYWORDS:

computational literary studies; scholarly corpora; reception studies; text reuse; Thomas Pynchon

TO CITE THIS ARTICLE:

Ketzan, E. (2026). The Open Pynchon Reception Corpus: From Bibliographic Metadata to Reusable Scholarly Full Text. *Journal of Open Humanities Data*, 12: 142, pp. 1–6. DOI: https://doi.org/10.5334/johd.632

(1) OVERVIEW

REPOSITORY LOCATION

The data supporting this article is openly available from the King's College London research data repository, KORDS, at https://doi.org/10.18742/33188055.

CONTEXT

This paper introduces the *Open Pynchon Reception Corpus* (OPRC), whose aim is to provide all known scholarly articles (and one monograph) devoted to the fiction of a single contemporary literary author, Thomas Pynchon, that are in English and available under a CC BY licence. The original PDFs are included in the corpus release, as well as new .txt transcriptions and a table of bibliographic metadata including license and DOI. The aims of the OPRC are to provide computational literary studies (CLS) with an open resource for reproducible studies centred around the quantitative analysis of literary reception.

The OPRC is a follow-up project to the *Thomas Pynchon Online Bibliography* (TPOB), a structured Zotero bibliography for resource discovery and bibliometric research, which includes metadata on 2,533 scholarly studies: 84 monographs, 1,234 articles, 915 chapters or sections, and 300 reviews of Pynchon studies (Ketzan, 2025). This new project, OPRC, releases the full-text subset of the TPOB that can be shared open access, as it is available under CC BY. OPRC thus moves from metadata to the texts of scholarly reception, while retaining links to the bibliographic records in TPOB.

CLS research has sought to enlarge the scale of literary-historical inquiry through distant reading (Moretti, 2013) and corpus-wide "macroanalysis" of linguistic and thematic patterns (Jockers, 2013). Literary scholarship itself has also been analysed as a corpus, as in Goldstone and Underwood's (2014) study of disciplinary change across journal articles. As CLS continues to grow, the importance of reproducibility in CLS and DH research widely has been stressed (Ries et al., 2023), and OPRC aims to contribute a dataset for reproducible pipelines by presenting a body of CC BY literary scholarship on a single authorial subject.

A rapidly growing body of CLS applies computational methods to reader reception, such as research questionnaires (van Dalen-Oskam, 2023), reader-generated reviews (e.g. Holur et al., 2021; Walsh & Antoniak, 2021; Feldkamp Moreira et al., 2025), digital social reading (Rebora and Pianzola, 2018; Rebora et al. 2021), and fanfiction (Pianzola et al. 2020). Computational reception studies involving professional criticism also has important digital precedents. The Dartmouth Dante Project assembled line-linked commentaries spanning centuries (Hollander, 1989); scholars have modelled expert critical readings of Fernando Pessoa's *Book of Disquiet* (Portela et al., 2024); and text re-use methods have been proposed to trace scholarly commentary on canonical English texts such as *Middlemarch* (Eckert & Terlunen, 2024) and German literary texts (Arnold & Jäschke, 2021, 2023). The OBVIL project provides access to a large TEI-encoded corpus of French literary criticism (Alexandre, 2019; OBVIL, n.d.), while *Reception Reader* supports exploration of textual reuse across early modern publications (Rosson et al., 2023). These resources vary in object and access; several are searchable environments rather than downloadable open-licensed corpora, and the *Middlemarch* study relies on JSTOR's Text Analysis Support service, which provides scholars with the OCR-ed full texts of scholarly outputs.[1] OPRC's narrower contribution is a downloadable, CC BY corpus of contemporary scholarly reception. It is intended for reproducible research, not as a comprehensive or representative sample of Pynchon studies.

The OPRC corpus release follows the FAIR principles (Wilkinson et al., 2016). A repository DOI, item DOIs, and stable corpus identifiers support findability. Open, documented formats, along with metadata support access and interoperability. Item-level provenance, licence evidence, versioning, split-package validation, and checksums support audit and reuse.

(2) METHOD

STEPS

Candidate scholarly texts were identified through TPOB and searches of open-access journal and publisher sites. Each source's licence was checked before inclusion, and text was extracted from the

1 https://support.jstor.org/hc/en-us/sections/32485095722135-JSTOR-Text-Analysis-Support.

DOI: 10.5334/johd.632

publisher PDF or HTML. Each document was converted to UTF-8 plain text and given a standard metadata header. Normalisation in .txt transformation removed page numbers and layout-driven line breaks while preserving paragraphs, headings, notes, and bibliographies. The sole CC BY scholarly monograph devoted to Pynchon, Hänggi's *Pynchon's Sound of Music* (2020), received additional deterministic cleaning of page markers, end-of-line hyphenation, and page layout.

The release metadata records the author, title, year, DOI, publisher URL, source and extraction method, normalisation, item-level licence and evidence, language, and word counts. Per-file SHA-256 checksums are included in the metadata along with a SHA256SUMS manifest, verifiable with a bundled validator, for full replicability.

The corpus release contains the Pynchon studies texts and CSV metadata, the resources that many common CLS analysis methods require. The companion provenance archive contains a wider set of code and documentation: the source PDFs, the saved publisher HTML in one case, JSON Lines metadata, the licence-evidence audit, and two Python scripts (the HTML extractor and a validator for the release).

SAMPLING STRATEGY

OPRC v1.0.0 includes substantive, English-language scholarship about Pynchon licensed under CC BY. Its 52 texts comprise 46 articles from *Orbit: A Journal of American Literature* (the *de facto* leading Pynchon studies journal), five articles from other journals, and Hänggi's monograph. The corpus includes publications from 2012 to 2025 and contains 597,608 word tokens excluding metadata.

Scholarly book reviews, announcements, and works not substantively about Pynchon were excluded. Works carrying CC NonCommercial, NoDerivatives, or ShareAlike conditions were also excluded, even when openly accessible. The reason is that although CC NC, ND, and SA are ostensibly "open" licenses, the NoDerivatives restriction in particular may not allow a critical step in my corpus's creation: the transformation of PDF to txt file via OCR. Although my aims are purely non-commercial, I also aimed to release the most unrestricted corpus, and thus exclude NonCommercial licensed texts, as well. Three open-access monographs engaging substantially with Pynchon, by Engelhardt (2018), Ketzan (2022), and Dehdarirad (2023), were thus excluded because their CC licences add NonCommercial and/or NoDerivatives restrictions. The CC BY criterion maximises legal clarity for redistribution but creates a strong publisher, date, and language bias; OPRC thus represents the reusable CC BY portion of Pynchon studies, not the entire field as described by TPOB.

These licensing issues highlight how little literary scholarship is currently available for truly open access re-use, even where authors and publishers embrace CC licenses generally. The fact that three CC-licensed monographs cannot be re-distributed in this project's open corpus format serves as evidence that CC BY should be the preferred licence for open access scholarship, if future computational processing is desired. The limitations of CC ND and NC licences are reflected in the *Journal of Open Humanities Data* submission policy, which recommends against these licences (*Journal of Open Humanities Data*, n.d.), and discussed by, for example, Vézina, who notes that articles published under an ND licence "are not considered OA" under leading open access frameworks (2020).

QUALITY CONTROL

The licence of each text in the corpus was manually checked. Twenty-six works explicitly state CC BY 4.0, Hänggi's monograph states CC BY 3.0 Switzerland, and 25 early *Orbit* PDFs state Creative Commons Attribution without specifying a version. Metadata was checked for unique corpus identifiers, paths, Zotero keys (from TPOB), and DOIs. The included validator script confirms the text-PDF pairs, verifies text and PDF hashes against the metadata, checks accepted licences, and verifies each archive's checksum manifest.

(3) DATASET DESCRIPTION

REPOSITORY NAME

King's Open Research Data System (KORDS)

OBJECT NAME

Open Pynchon Reception Corpus (OPRC), version 1.0.0

FORMAT NAMES AND VERSIONS

UTF-8 plain text (.txt); CSV; JSON Lines; PDF; HTML; Markdown; Python 3; YAML/Citation File Format; SHA-256 checksum manifest

CREATION DATES

2026-06-16 to 2026-08-08

DATASET CREATOR

Erik Ketzan, Department of Digital Humanities, King's College London (Conceptualization, Data curation, Investigation, Methodology, Software, Validation)

LANGUAGE

English

LICENCE

Collected corpus texts and normalised texts retain their recorded item-level CC BY licences. Original corpus selection, arrangement, metadata, and documentation are CC BY 4.0; Python scripts are MIT licensed. Embedded third-party material within the scholarly texts, such as quotations, remains under its own copyright and is not relicensed.

PUBLICATION DATE

2026-08-10

(4) REUSE POTENTIAL

OPRC supports both computational and qualitative study of contemporary literary reception, and any other uses permitted by CC BY. Possible uses include a variety of CLS, NLP, and DH studies, including computational reception studies and citation-network enrichment. OPRC enables the identification and mapping of quotations, allusions, and repeated passages between Pynchon's fiction and its scholarly reception (Ketzan and Torde, forthcoming). Because metadata, provenance, licence evidence, processing code, and source files are packaged together, users can audit individual documents, or, as OPRC is a rare example of a contemporary literary reception corpus, potentially use the dataset in teaching open corpus construction.

The principal limitation of OPRC is that it can only include CC BY material, which represents a fraction of extant Pynchon studies. The corpus is small relative to TPOB's 2,500+ listed sources, is dominated by one journal (*Orbit*), limited to English, and concentrated in 2012–2025. The restriction to English-language scholarship is a pragmatic first step, and future work involving non-English languages should be considered. Further, the creation and distribution of the OPRC make no claims about the academic quality of the contents. Plain-text conversion removes page layout and may retain occasional source or extraction artefacts (which is why the original PDFs are also provided). Finally, the CC BY status of any scholarly work does not extend to quotations from copyright-protected literary and scholarly texts quoted within them; re-users remain responsible for assessing third-party material in their own jurisdiction and use case. I hope that OPRC may support future studies on this specific author, Pynchon, and CLS studies more broadly.

AI DECLARATION

Claude Code and OpenAI Codex were used as assistive software-development tools, assisting with the creation of Python scripts for deterministic text normalisation, metadata construction, validation, and packaging. The author determined all inclusion, licensing, and editorial decisions, and reviewed the code and transformed outputs.

DATA ACCESSIBILITY STATEMENT

The data supporting this article is openly available from the King's College London research data repository, KORDS, at https://doi.org/10.18742/33188055. The repository record provides two archives: a data archive containing the corpus texts and metadata, and a companion provenance archive containing the original PDFs, publisher HTML in one case, extraction and validation scripts, and checksum manifests.

ACKNOWLEDGEMENTS

The author thanks the organizers and audience of International Pynchon Week 2026 at TU Dortmund University for valuable feedback.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Erik Ketzan: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

AUTHOR AFFILIATIONS

Erik Ketzan orcid.org/0000-0003-0264-2157
Department of Digital Humanities, King's College London, London, United Kingdom

REFERENCES

Alexandre, D. (2019). Les bibliothèques numériques du labex OBVIL, un patrimoine? In É. Anheim, A.-J. Etter, G. Glasson Deschaumes, & P. Liévaux (Eds.), *Les patrimoines en recherche(s) d'avenir* (pp. 175–186). Presses universitaires de Paris Nanterre. https://doi.org/10.4000/books.pupo.22980

Arnold, F., & **Jäschke, R.** (2021). Lotte and Annette: A framework for finding and exploring key passages in literary works. In M. Hämäläinen, K. Alnajjar, N. Partanen, & J. Rueter (Eds.), *Proceedings of the Workshop on Natural Language Processing for Digital Humanities* (pp. 55–63). NLP Association of India. https://aclanthology.org/2021.nlp4dh-1.7/

Arnold, F., & **Jäschke, R.** (2023). A novel approach for identification and linking of short quotations in scholarly texts and literary works. *Journal of Computational Literary Studies*, *2*(1). https://doi.org/10.48694/jcls.3590

Dehdarirad, A. (2023). *"From faraway California": Thomas Pynchon's aesthetics of space in the California trilogy*. Sapienza Università Editrice. https://doi.org/10.13133/9788893772877

Eckert, S., & **Terlunen, M.** (2024). What we quote: Disciplinary history and the textual atmospheres of *Middlemarch*. *Victorian Studies*, *66*(4), 557–584. https://doi.org/10.2979/vic.00193

Engelhardt, N. (2018). *Modernism, fiction and mathematics*. Edinburgh University Press. https://doi.org/10.3366/edinburgh/9781474416238.001.0001

Feldkamp Moreira, P., Bizzoni, Y., Jacobsen, M., Thomsen, M. R., & **Nielbo, K. L.** (2025). The Goodreads' ›Mediocre‹: Assessing a Grey Area of Literary Judgements. In *Weder Fail noch Lobgesang. Nichteindeutige Wertung von Literatur im digitalen Raum: Sonderband 6* (Version 1.0). Forschungsverbund Marbach Weimar Wolfenbüttel/Digital Humanities im deutschsprachigen Raum e. V. https://doi.org/10.17175/SB006_002

Goldstone, A., & **Underwood, T.** (2014). The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, *45*(3), 359–384. https://doi.org/10.1353/nlh.2014.0025

Hänggi, C. (2020). *Pynchon's sound of music*. Diaphanes. https://doi.org/10.4472/9783035802337

Hollander, R. (1989). The Dartmouth Dante Project. *Quaderni d'Italianistica*, *10*(1–2), 287–298. https://doi.org/10.33137/q.i..v10i1-2.10436

Holur, P., Shahsavari, S., Ebrahimzadeh, E., Tangherlini, T. R., & **Roychowdhury, V.** (2021). Modelling social readers: Novel tools for addressing reception from online book reviews. *Royal Society Open Science*, *8*(12), 210797. https://doi.org/10.1098/rsos.210797

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press. https://doi.org/10.5406/illinois/9780252037528.001.0001

Journal of Open Humanities Data. (n.d.). *Which open license should I apply to my data?* https://openhumanitiesdata.metajnl.com/about#which-open-license-should-i-apply-to-my-data

Ketzan, E. (2022). *Thomas Pynchon and the digital humanities: computational approaches to style.* Bloomsbury Academic. https://doi.org/10.5040/9781350211865

Ketzan, E. (2025). A bibliometric overview of Pynchon studies: introducing the Thomas Pynchon Online Bibliography (TPOB). *Orbit: A Journal of American Literature*, *12*(1), 1–25. https://doi.org/10.16995/orbit.23348

Ketzan, E., & **Torde, I.** (forthcoming). *Not every match is a quotation: evaluating LLMs for quotation detection in literary reception.*

Moretti, F. (2013). *Distant reading.* Verso.

OBVIL. (n.d.). *Corpus critique* [TEI XML corpus]. Retrieved July 17, 2026, from https://obvil.github.io/critique/

Pianzola, F., Acerbi, A., & **Rebora, S.** (2020). Cultural accumulation and improvement in online fan fiction. In F. Karsdorp, B. McGillivray, A. Nerghes, & M. Wevers (Eds.), *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)* (pp. 2–11). CEUR-WS. https://ceur-ws.org/Vol-2723/short8.pdf

Portela, M., Marques, A., & **Pereira, L. L.** (2024). Modelling and visualising critical reading through an intertextual interface. *Umanistica Digitale*, *8*(17), 121–142. https://doi.org/10.6092/issn.2532-8816/19462

Rebora, S., Boot, P., Pianzola, F., Gasser, B., Herrmann, J. B., Kraxenberger, M., Kuijpers, M. M., Lauer, G., Lendvai, P., Messerli, T. C., & **Sorrentino, P.** (2021). Digital humanities and digital social reading. *Digital Scholarship in the Humanities*, *36*(Supplement_2), ii230–ii250. https://doi.org/10.1093/llc/fqab020

Rebora, S., & **Pianzola, F.** (2018). A New Research Programme for Reading Research: Analysing Comments in the Margins on Wattpad. *DigitCult – Scientific Journal on Digital Cultures*, *3*(2), 19–36. https://doi.org/10.54103/2531-5994/31142

Ries, T., van Dalen-Oskam, K., & **Offert, F.** (2023). Reproducibility and explainability in digital humanities. *International Journal of Digital Humanities*, *5*(2–3), 247–251. https://doi.org/10.1007/s42803-023-00078-7

Rosson, D., Mäkelä, E., Vaara, V., Mahadevan, A., Ryan, Y., & **Tolonen, M.** (2023). Reception Reader: Exploring text reuse in early modern British publications. *Journal of Open Humanities Data*, *9*, Article 5. https://doi.org/10.5334/johd.101

van Dalen-Oskam, K. (2023). *The Riddle of Literary Quality: A Computational Approach.* Amsterdam University Press. https://doi.org/10.5117/9789048558148

Vézina, B. (2020, April 21). *Why sharing academic publications under "No Derivatives" licenses is misguided.* Creative Commons. Retrieved from source https://creativecommons.org/2020/04/21/academic-publications-under-no-derivatives-licenses-is-misguided/ (last accessed: September 16, 2026).

Walsh, M., & **Antoniak, M.** (2021). The Goodreads "classics": A computational study of readers, Amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics*, *6*(2). https://doi.org/10.22148/001c.22221

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 160018. https://doi.org/10.1038/sdata.2016.18

TO CITE THIS ARTICLE:
Ketzan, E. (2026). The Open Pynchon Reception Corpus: From Bibliographic Metadata to Reusable Scholarly Full Text. *Journal of Open Humanities Data*, 12: 142, pp. 1–6. DOI: https://doi.org/10.5334/johd.632

Submitted: 10 August 2026
Accepted: 30 September 2026
Published: 02 November 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.