Journal of Open Humanities Data

Hors d'usage: Curated French Fiction from Gallica, 1800–1950

DATA PAPER

JEAN BARRÉ

ubiquity press

ABSTRACT

This paper presents a curated, ready-to-use corpus of 10,898 French novels published between 1800 and 1950, derived from the *Fictions littéraires de Gallica* collection. The corpus is designed as an infrastructure for the empirical study of nineteenth-century popular fiction — the literature *hors d'usage* that once constituted the bulk of novelistic production but has since fallen out of circulation and out of literary memory. A reproducible pipeline filters low-quality OCR scans, removes editorial duplicates through a hybrid procedure combining MinHash near-duplicate detection and rule-based metadata matching, and enriches each record with normalized author identities and BnF/Wikidata-cross-referenced biographical metadata. Two genre labels — adventure (1,428 novels) and detective fiction (830 novels) — are provided as flags in the metadata table, illustrating one form of downstream reuse. The dataset, the pipeline, the per-novel decisions, and the audit logs are openly released.

CORRESPONDING AUTHOR:
Jean Barré
Lattice, CNRS, ENS-PSL, Paris, France
jean.barre@ens.psl.eu

KEYWORDS:
French literature; popular fiction; Gallica; corpus curation; OCR; computational literary studies

TO CITE THIS ARTICLE:
Barré, J. (2026). *Hors d'usage*: Curated French Fiction from Gallica, 1800–1950. *Journal of Open Humanities Data*, 12: 116. DOI: https://doi.org/10.5334/johd.597

1 OVERVIEW

REPOSITORY LOCATION

The dataset is openly archived on Zenodo at https://doi.org/10.5281/zenodo.20233178. The deduplication and enrichment pipeline, with documentation, is mirrored on GitHub and archived for long-term availability on Zenodo at https://doi.org/10.5281/zenodo.21508679. The deposit is versioned: version 2 (21 May 2026) adds the genre-annotation columns accidentally omitted from the initial upload, all other components being unchanged, and future versions may address issues reported by users, with all changes documented in the Zenodo version history.

CONTEXT

The corpus was produced as part of doctoral research on the computational analysis of French popular fiction. Its broader motivation is to make empirically accessible what Cohen (1999) calls literature *hors d'usage*: works fallen out of circulation and out of literary memory. We use the term in a production-side sense: lacking sales figures or print runs, the dataset does not measure popularity, but it captures the breadth of what was printed. Derived from the BnF's legal-deposit holdings — an estimated 40% of nineteenth-century novelistic production (Langlais, 2019) — the corpus approximates the mass of published fiction rather than a retrospectively selected canon: production peaks in the 1880s, the golden age of the *roman-feuilleton* (Figure 1), and most of its 3,632 authors are absent from literary histories or anthologies.

The dataset is built on top of the upstream *Fictions littéraires de Gallica* collection assembled by Langlais (2021), and continues a line of work that recovers genre and stylistic categories from full texts rather than catalog metadata (Langlais, 2019; Underwood, 2019). Methodologically, our deduplication pipeline draws on the hybrid identifier-plus-fuzzy strategy recently used for shadow-library cataloguing at scale (Mazières & Poibeau, 2025), adapted here to the specifics of OCR-derived French sources. The resulting corpus extends and refines an earlier 12,176-novel version first used in Barré (2024b) to study intertextual structure, and has subsequently supported published research on adventure-novel retrieval (Barré, 2026) and passage-level genre detection (Barré, 2024a). Within the landscape of fiction corpora for literary history, the dataset complements English-language resources such as the NovelTM datasets (Underwood et al., 2020), and the multilingual ELTeC collections (Schöch et al., 2021); for French, it stands alongside the *Chapitres* corpus (Leblond, 2022) and CIDRE (Seminck & Gambette, 2022), and — for the immediately preceding period — the MiMoText collection of eighteenth-century French novels (Röttgermann, 2024), from which it differs in scale, in its focus on non-canonical production, and in the release of the curation decisions themselves.

2 METHOD

STEPS

The starting point is the *Fictions littéraires de Gallica* collection compiled by Langlais (2021), from which we draw the subset that has been filtered for OCR quality (95% correct-word ratio against a manually corrected French lexicon), with *Œuvres complètes* volumes removed — they aggregate distinct novels under a single record and distort similarity computations through size asymmetry — and with a first pass of rule-based metadata deduplication on exact author–title matches. This yields a working corpus of 12,176 candidate records, on which the present pipeline applies four further stages: a temporal restriction, a content-aware deduplication, a metadata enrichment, and a genre-labeling stage.

The temporal restriction discards records published before 1800, where the upstream metadata is noisier and the OCR less reliable; this removes 267 records and yields 11,909 candidates in the 1800–1950 window.

The deduplication stage combines content-level similarity with a second, deeper pass of rule-based metadata matching, with each per-pair decision logged for audit. On the content side, a 256-permutation MinHash signature is computed for every text after Gallica-specific cleaning: suppression of paratextual lines, page numbers, image captions, and a 5% trim on each end

to neutralize prefaces and editorial colophons. Pairs with Jaccard similarity at or above 0.55 are surfaced through Locality-Sensitive Hashing (Leskovec et al., 2020). On the metadata side, four cascading passes are applied within each author group: an EXACT pass on normalized titles; a RACINES pass on strict prefixes (typically a bare title versus the same with an editorial suffix such as *par M. Dumas*); a RACINES_2 pass on shared two-token roots, with risk tags for generic roots and oversized clusters; and a FUZZY pass using token-sort similarity above 0.85. A pair is automatically merged only if (i) the metadata pass returns an EXACT or long-prefix match, or (ii) the MinHash similarity exceeds 0.75 *and* the two records share an author. Edge cases are routed to a manual-review queue. This conservative design proceeds in two passes: a hybrid pass combining content and metadata signals removes 917 duplicates, yielding an intermediate corpus of 10,992 novels, after which a second pass merges 94 remaining same-author, title-compatible editions that survived the first pass. The two passes remove 1,011 duplicates in total, yielding 10,898 distinct novels. When a pair is merged, the retained witness is the earliest publication date. Because the procedure operates on textual similarity rather than on bibliographic identity, substantially rewritten editions of the same work fall below the similarity thresholds and are retained as distinct texts, while lightly revised editions are collapsed onto a single witness; the retained text therefore carries no philological claim about which state of the work it represents.

The metadata enrichment stage corrects OCR artefacts in titles and author names, normalizes author identities across pseudonyms and orthographic variants, and recovers biographical dates through a cross-reference between BnF authority records and Wikidata. The final corpus contains 10,898 novels by 3,632 distinct authors, with 86% of records linked to a BnF authority identifier and 85% to a Wikidata entity. Its temporal distribution is shown in Figure 1.

The final stage adds two genre labels to the metadata table: an `is_adventure` flag and an `is_detective` flag, produced by supervised classification anchored in historically situated seed lists and refined over a *k*-nearest-neighbor similarity graph (Barré, 2026). The labels are introduced here because they are part of the released resource and of its quality control; their construction and reuse potential are detailed in Section 4.

SAMPLING STRATEGY

The corpus inherits the sampling regime of the upstream Gallica collection, which itself reflects the French legal-deposit system: Langlais (2019) estimates that roughly 40% of nineteenth-century French novels are covered. Post-1950 coverage drops sharply due to copyright restrictions on full-text release.

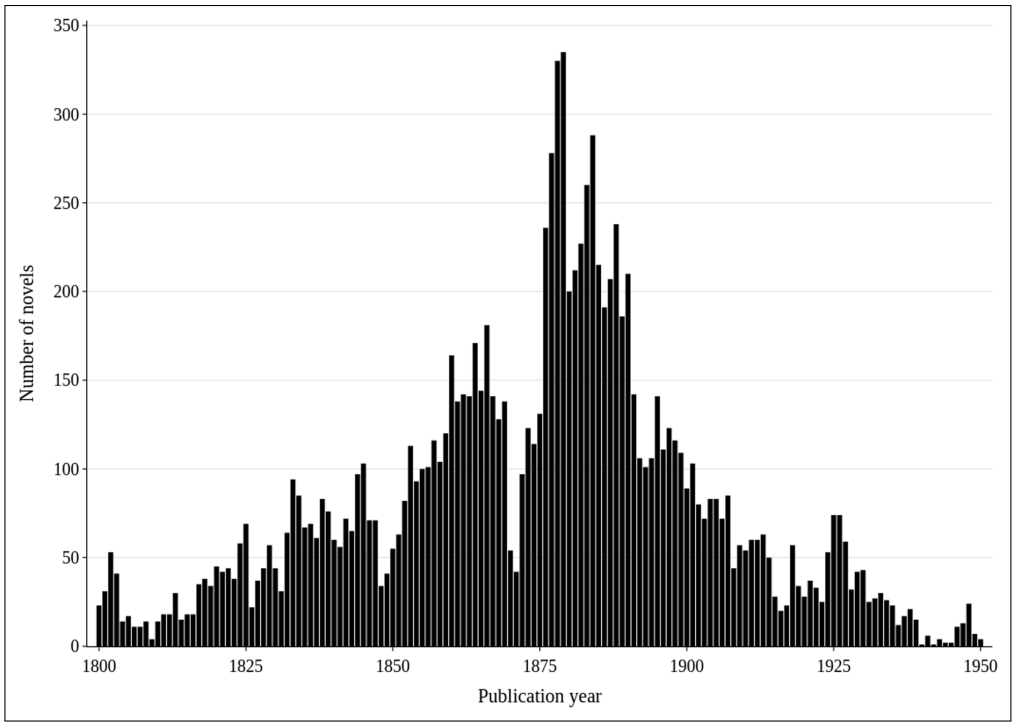

Figure 1 Temporal distribution of the 10,898 novels. Production peaks in the 1880s, reflecting the golden age of the *roman-feuilleton*. The drop after 1900 is partly an artefact of Gallica's digitization coverage and of copyright constraints on twentieth-century works.

QUALITY CONTROL

Three control mechanisms were applied in addition to the upstream OCR filter described above. The deduplication procedure was validated against a manually annotated sample of 82 pair candidates, stratified across eight Jaccard buckets between 0.50 and 1.00; the operational auto-merge threshold was set so that the lower bound of the 95% bootstrap confidence interval on precision exceeds the target value (see Appendix 4, Figure A.1). Every per-pair decision is logged in the released `decisions.tsv`, allowing full reverse-engineering of the curated corpus. Finally, the adventure label was validated externally: the *Chapitres* corpus (Leblond, 2022), a collection of French novels with independently curated genre metadata, provides 364 novels labeled as adventure fiction, of which our classifier retrieves 81% (Barré, 2026).

3 DATASET DESCRIPTION

REPOSITORY NAME

Zenodo (primary archival deposit, versioned); GitHub (working mirror, pipeline code).

OBJECT NAME

`txts_gallica.tar.gz` (full-text archive) and `metadata_gallica.tsv` (metadata table).

FORMAT NAMES AND VERSIONS

A gzipped tar archive containing UTF-8 plain-text files (`.txt`), one per novel; an accompanying tab-separated values file (`.tsv`) with per-novel metadata; documentation in Markdown.

CREATION DATES

2024-06-01 to 2026-05-17.

DATASET CREATORS

Jean Barré (Lattice, CNRS / ENS-PSL). The dataset reuses, as an upstream resource, the Gallica digital collections of the Bibliothèque nationale de France and the *Fictions littéraires de Gallica* collection compiled by Pierre-Carl Langlais (Langlais, 2021).

LANGUAGE

French (text content, titles); English (variable names, documentation).

LICENSE

The full texts originate from Gallica and are in the public domain in France. In agreement with the Bibliothèque nationale de France, any reuse of the textual files must include the attribution “Source gallica.bnf.fr / Bibliothèque nationale de France” and a reference to the Gallica conditions of use.[1] The curated metadata, deduplication pipeline, and derived files are distributed under CC BY 4.0.

PUBLICATION DATE

2026-05-17.

4 REUSE POTENTIAL

The dataset is designed as an infrastructure layer for the computational study of French fiction, and is to our knowledge the largest curated and quality-controlled corpus available for that purpose.

GENRE LABELS AS WORKED EXAMPLES

Two genre annotations are distributed as columns in the metadata table. The first flags 1,428 adventure novels, retrieved through supervised classification anchored in a historically

1 https://gallica.bnf.fr/accueil/fr/html/conditions-dutilisation-de-gallica.

situated seed list, followed by graph-based refinement on a *k*-nearest-neighbor similarity graph computed in a fine-tuned embedding space (Barré, 2026). The second flags 830 detective novels, produced through the same pipeline with a different seed list. Both are released with per-novel classifier probabilities, so that users can construct working corpora at alternative confidence thresholds, or apply the methodology to other genres. They illustrate that the dataset is structured to support genre-level work: full texts are recoverable, author identities are stable, and per-novel metadata is rich enough to anchor empirical comparisons.

STYLISTIC AND BIBLIOMETRIC STUDIES

The size and temporal range of the corpus enable analyses of lexical and rhetorical change across more than a century of French prose fiction. The author distribution is sharply long-tailed: a handful of canonical or commercially dominant signatures account for a substantial volume share, while most of the 3,632 authors appear with only one or two titles. The corpus is thus well-suited to studies of authorial trajectories and of the canon/archive distinction (Moretti, 2000; Algee-Hewitt et al., 2016; Barré et al., 2023): debates on canonicity and the "great unread" (Cohen, 1999) can be tested at scale on the French fiction that originally prompted the category.

LITERARY HISTORY AND PEDAGOGY

The dataset enables empirical reconsideration of nineteenth-century categories displaced from contemporary curricula, and provides a manageable but realistic resource for teaching corpus curation, stylometry, supervised classification, embedding-based retrieval, and reproducibility workflows.

LIMITATIONS

Coverage is bounded upstream by Gallica's digitization choices and by copyright. OCR noise persists despite the 95% filter and may affect lexical-level analyses. The two genre labels operationalize specific literary-historical perspectives and should not be treated as exhaustive.

APPENDIX A: DEDUPLICATION VALIDATION

A stratified sample of 82 candidate duplicate pairs was manually annotated, drawn across eight Jaccard buckets between 0.50 and 1.00. Each pair was labeled `YES` (same work, possibly different editions), `NO` (distinct works), or `UNSURE`; 78 of the 82 received a conclusive label. Precision and recall were estimated at each candidate threshold $T \in [0.50, 1.00]$ via the empirical proportions among annotated pairs, with 95% confidence intervals obtained through

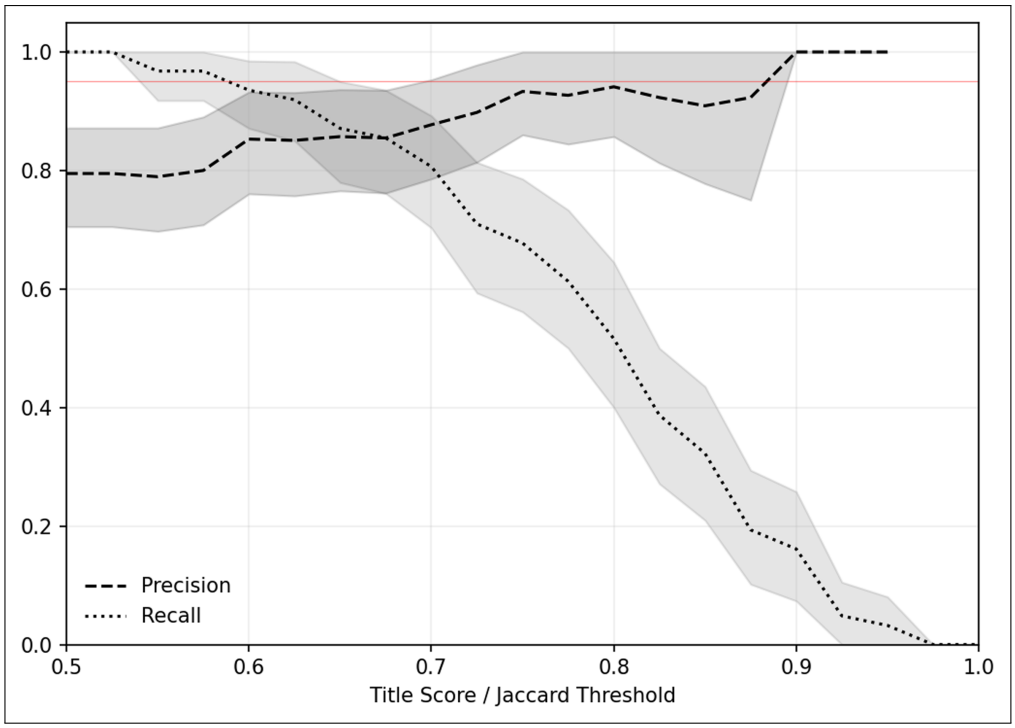

Figure A.1 Precision (dashed) and recall (dotted) of MinHash-based duplicate detection as a function of the Jaccard threshold, with 95% bootstrap confidence intervals (shaded). The hybrid pipeline used for the released corpus combines this content-level signal with the rule-based metadata signal described in Section 2, which is why the operational regime sits at a precision level higher than the curve indicates for MinHash alone.

1,000 bootstrap resamples. The operational threshold was selected as the smallest T at which the lower bound of the 95% precision CI reaches the target value; the released corpus combines a MinHash threshold of 0.75 with author-coherence constraints from the metadata pass.

DATA ACCESSIBILITY STATEMENT

Dataset: Zenodo, https://doi.org/10.5281/zenodo.20233178. Pipeline: GitHub and Zenodo, https://doi.org/10.5281/zenodo.21508679.

ACKNOWLEDGEMENTS

The author thanks the Bibliothèque nationale de France for the open Gallica program, and in particular Jean-Philippe Moreux and the BnF DataLab team — notably Marie Carlin — for ongoing support to scholarly reuse. Thanks also to Pierre-Carl Langlais for the upstream *Fictions littéraires de Gallica* collection, and to Aude Leblond and colleagues for the *Chapitres* validation corpus.

FUNDING STATEMENT

This work benefited from funding from the French State, managed by the Agence Nationale de la Recherche, under the France 2030 program – PRAIRIE-PSAI (Paris School of Artificial Intelligence, grant reference ANR-23-IACL-0008) and the Major Research Program of PSL Research University “CultureLab” (grant reference ANR-10-IDEX-0001).

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Jean Barré: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

AUTHOR AFFILIATIONS

Jean Barré orcid.org/0000-0002-1579-0610
Lattice, CNRS, ENS-PSL, Paris, France

REFERENCES

Algee-Hewitt, M., Allison, S., Gemma, M., Heuser, R., Walser, H., & **Moretti, F.** (2016). Canon/Archive. Large-scale Dynamics in the Literary Field. *Pamphlets of the Stanford Literary Lab*, (11). https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf

Barré, J. (2024a). Détection automatique de l’architextualité dans le roman d’aventures (1870–1930). In *Humanistica 2024*. Association francophone des humanités numériques. https://hal.science/hal-04559749

Barré, J. (2024b). Latent Structures of Intertextuality in French Fiction. In *Proceedings of the Computational Humanities Research Conference 2024* (pp. 21–26). https://ceur-ws.org/Vol-3834/paper97.pdf

Barré, J. (2026). In Search of Lost Adventure Novels: Supervised Genre Retrieval and Corpus Refinement in Gallica. In S. Hamilton, E. Öhman, R. M. M. Hicke, Y. Bizzoni, A. Bax, J. A. Matthews, & M. Hämäläinen (Eds.), *Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities* (pp. 255–263). San Association for Computational Linguistics. https://aclanthology.org/2026.nlp4dh-1.24/

Barré, J., Camps, J.-B., & **Poibeau, T.** (2023). Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature. *Journal of Cultural Analytics*, *8*(3). https://doi.org/10.22148/001c.88113

Cohen, M. (1999). *The sentimental education of the novel*. Princeton University press. https://doi.org/10.1515/9780691188249

Langlais, P.-C. (2019). *Reconstituer les genres romanesque sur Gallica: essai de classification automatisée de 1500 romans (1815-1850)* [Billet]. https://scoms.hypotheses.org/986

Langlais, P.-C. (2021). *Fictions littéraires de Gallica.* Zenodo. https://doi.org/10.5281/zenodo.4751204

Leblond, A. (2022). *Corpus Chapitres.* Zenodo. https://doi.org/10.5281/zenodo.7446728

Leskovec, J., Rajaraman, A., & **Ullman, J. D.** (2020). *Mining of Massive Datasets*. Cambridge University Press, 3 edn. https://doi.org/10.1017/9781108684163

Mazières, A., & **Poibeau, T.** (2025). MajinBook: An open catalogue of digitally mediated world literature. https://arxiv.org/abs/2511.11412

Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, (1), 54–68. https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature

Röttgermann, J. (2024). The Collection of Eighteenth-Century French Novels 1751–1800. *Journal of Open Humanities Data*, *10*, 31. https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.201/

Schöch, C., Patras, R., Erjavec, T., & **Santos, D.** (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, *1*, 25. https://www.modernlanguagesopen.org/articles/10.3828/mlo.v0i0.364/

Seminck, O., & **Gambette, P.** (2022). *CIDRE: Corpus for Idiolectal Research.* https://doi.org/10.82270/cidre/v3 (ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr)

Underwood, T. (2019). *Distant horizons: digital evidence and literary change*. The University of Chicago Press. https://doi.org/10.7208/chicago/9780226612973.001.0001

Underwood, T., Kimutis, P., & **Witte, J.** (2020). NovelTM Datasets for English-Language Fiction, 1700-2009. *Journal of Cultural Analytics*. https://culturalanalytics.org/article/13147-noveltm-datasets-for-english-language-fiction-1700-2009

TO CITE THIS ARTICLE:
Barré, J. (2026). *Hors d'usage*: Curated French Fiction from Gallica, 1800–1950. *Journal of Open Humanities Data*, 12: 116. DOI: https://doi.org/10.5334/johd.597

Submitted: 26 May 2026
Accepted: 27 July 2026
Published: 20 August 2026

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