theories and methodologies

# Slow Metadata

RACHEL SAGNER BUURMA AND JON SHAW

THE BIBLIOGRAPHIC RECORDS IN LIBRARIES' SEARCHABLE ONLINE PUBLIC ACCESS CATALOGS (OPAC) HAVE RECENTLY TAKEN ON A NEW ROLE as a source of bibliographic data that can be aggregated, shared, circulated, manipulated, transformed, studied, and interpreted.[1] Scholars' new awareness of library catalogs not just as aids to locating books and other materials but as sources of bibliographic information that researchers can manipulate and transform has inspired new scholarship on the history of the catalog and a new focus on how the catalog, in both its analog and digital forms, shapes bibliographic knowledge. Our Early Novels Dataset (END) project, for example, uses methods from book history, library science, and literary studies to think about the shape and history of the bibliographic metadata in the library catalog. Our research group's collective experiments with bibliographic metadata ask what happens when we look at the library catalog record not just as a utilitarian aid for searching or as an object of critique, but also as a work in progress with a literary character of its own. We ask what we can learn from the shape given to bibliographic information by the earlier catalogers whose records our project inherited and on whose expertise we draw. We also ask how the familiar languages of the library catalog record and the controlled bibliographic description might help make new forms of knowledge about books. And we press on the inevitable and generative tension between the particular perspective of the library catalogers who transform specific copies of physical books into bibliographic data and the informational fields dictated by machine-readable cataloging (MARC) descriptive standards.[2]

Our project asks in particular about the forms in which eighteenth-century books offered their readers and users their own bibliographic taxonomies and forms of search or access—and we attempt to represent these eighteenth-century taxonomies using the standard MARC fields of the online catalog record.[3] Each summer, we gather researchers together to describe a few dozen or a few hundred books in the collections of eighteenth-century works of fic-

RACHEL SAGNER BUURMA is associate professor of English literature at Swarthmore College. Her book *The Teaching Archive: A New History for Literary Study*, written with Laura Heffernan, will be published by the University of Chicago Press in the fall of 2020.

JON SHAW is associate vice provost and deputy university librarian at the University of Pennsylvania.

© 2020 RACHEL SAGNER BUURMA AND JON SHAW
*PMLA* 135.1 (2020), published by the Modern Language Association of America



https://doi.org/10.1632/pmla.2020.135.1.188 Published online by Cambridge University Press

tion in English that make up our case study.[4] These range from the canonical to the almost unknown, from the later-eighteenth-century books that announce themselves as novels to the surprising lives and adventures, collections of letters, dialogues with the dead, travelogues, and secret memoirs that populate the nascent field of long fictional narrative earlier in the century. Our undergraduate researchers examine copies of books and describe them in MARCXML-encoded bibliographic records using the controlled, MARC-based vocabulary we have developed. These student researchers learn about early novels, bibliographic description, and cataloging protocols from our team of professional catalogers, rare-book curators, digital scholarship specialists, and scholars of eighteenth-century literature. Each day of our summer program, our students spend the morning reading and describing; in the afternoon, they turn to individual projects, in which they create their own interpretations of our datasets.[5] We aim to foster an educational environment that encourages slow, careful, and personal bibliographic description, helping us pay close attention to the books yet also modeling the sociality of scholarship and of library work.

Our project draws on the conventions of the library catalog to describe the complicated paratexts of eighteenth-century novels—their text-heavy title pages, long tables of contents, elaborate prefaces, dedications, introductions, footnotes, and representations of authorship—as well as copy-specific information about marginalia, inscriptions, and bookplates.[6] Paratexts were crucial early modern reading aids—the interfaces and metadata of the eighteenth-century book. They described the book for readers and revealed the networks of relationships through which books were created and circulated. Eighteenth-century novels' paratexts allowed early readers to sort books based on topic, find particular information, and locate a work in its genre landscape; that is, they enabled forms of what we now call search. For example, the earliest readers of *The Life and Strange Surprising Adventures of Robinson Crusoe* (1719) encountered that book primarily as a story that fell into the categories of "Life" and "Adventure." *Pamela*'s status as a domestic novel, signaled in part by its heroine's first name on the title page and its epistolary form, was important to readers of its many editions.[7] Encoding information about paratexts and copy-specific details in controlled vocabularies makes it possible to imagine new ways of aggregating, collecting, and assembling early fiction according to some of the informational forms that earlier readers enjoyed. Novels, our project argues, are an important part of information history, and the history of metadata includes the title pages of eighteenth-century books.[8]

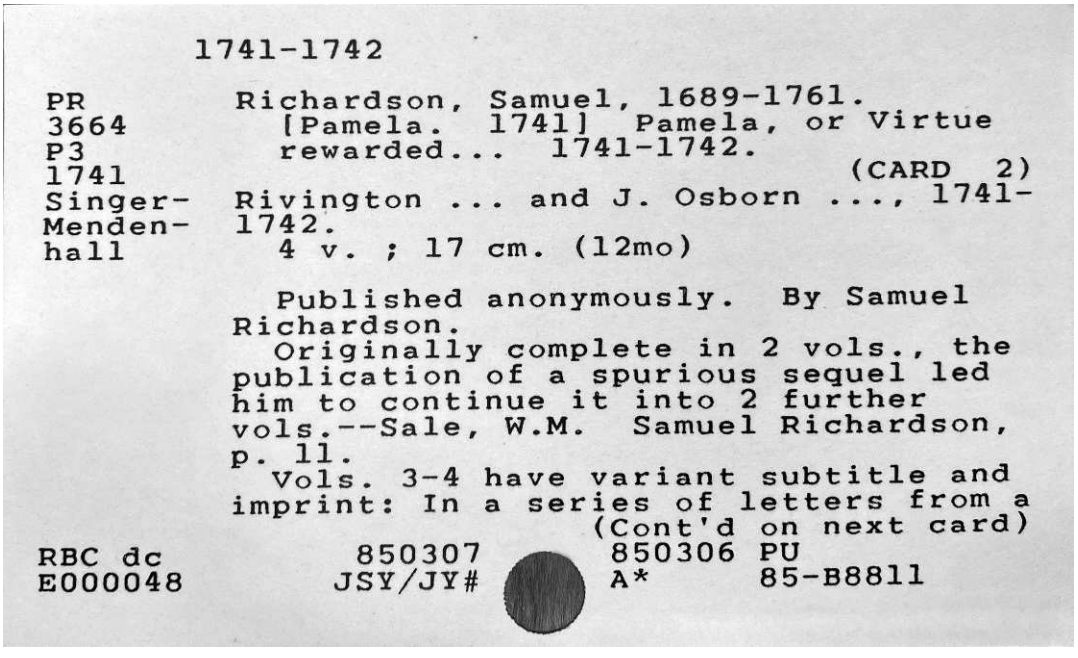
At the heart of the project is a bibliographic poetics based on the productive encounter between the irreducible, specific details of a copy of a book and the nuanced but standardizing languages of bibliographic description, library cataloging standards, MARCXML encoding, and tab-separated-value (TSV) files. Rather than sideline the encounter of eighteenth-century book technology with twentieth-century library cataloging standards and twenty-first-century data forms, our project centers it in both theory and practice. We rely on existing MARC standards but have created a custom ontology structure in MARC to capture bibliographic details of genre not usually included in a library catalog record.

Our project literally builds on the long histories of bibliographic description and the work of library catalogers; we begin describing a book by working from the book itself and a copy of its existing bibliographic record, usually one that has been created by a rare-book cataloger. Preserving the original record, we enrich it with our own categories. For example, to the 245 "title statement" MARC field, which almost always exists in a catalog record, we add the rarely used 246

https://doi.org/10.1632/pmla.2020.135.1.188 Published online by Cambridge University Press

MARC field in order to capture the full title in each of its variants across all volumes of the work.[9] The original cataloger's record therefore remains preserved and discernible in our record, its relation to our additions clear and the building blocks of the historic record visible. And despite consisting of highly controlled categories designed to standardize and aggregate bibliographic information, MARC also provides space for the cataloger's own perspective, a space our project embraces and even stretches beyond its usual purpose. We use the 500 field, a standard MARC designation for a general notes field in which discursivity is not usually encouraged, to capture extensive subjective or additional impressions of a book or of the process of cataloging it, while also preserving the long notes some previous catalogers have created.

We therefore answer Johanna Drucker's call for rethinking "data" as "capta" ("Humanities Approaches") and for creating methods for "generating capta that have some of the characteristics of humanities documents and expressions" ("Graphical Approaches" 248).[10] In some sense, of course, all bibliographic data—in fact, all data—are human-made and have the characteristics of humanities documents and expressions. A dataset reflects the specific forms, histories, and people that make it; it is a representation of its sources, and therefore requires humanities theories of representation and mediation to understand it. Using our bibliographic data as an example, we seek to demonstrate how bibliographic metadata will always reduce the books it describes, often in useful ways; as Katie Rawson and Trevor Muñoz note, "Cataloging's aim is to find a way to make items at least interchangeable enough not to break the system" (285). What we get in return for the reductions catalogers practice is a view of the multiplicity of characteristics—material, paratextual, generic—that early works of fiction share with one another.[11] And yet we continue to seek, as we do in the project's use of 500 notes, ways for catalogers to capture what Rawson and Muñoz, following Anna Lowenhaupt Tsing, call the "unscalable" elements of the dataset's sources, those elements of books that cannot be captured within the standards of MARC encoding (283; see Tsing 505).



1741-1742

PR
3664
P3
1741
Singer-
Menden-
hall

Richardson, Samuel, 1689-1761.
[Pamela. 1741] Pamela, or Virtue
rewarded... 1741-1742.
(CARD 2)
Rivington ... and J. Osborn ..., 1741-
1742.
4 v. ; 17 cm. (12mo)

Published anonymously. By Samuel
Richardson.
Originally complete in 2 vols., the
publication of a spurious sequel led
him to continue it into 2 further
vols.--Sale, W.M. Samuel Richardson,
p. 11.
Vols. 3-4 have variant subtitle and
imprint: In a series of letters from a
(Cont'd on next card)

RBC dc
E000048
850307
JSY/JY#
850306 PU
A*
85-B8811

FIG. 1
Card describing Richardson's *Pamela* (1741), from the University of Pennsylvania Libraries' Kislak Center for Special Collections, Rare Books, and Manuscripts.

https://doi.org/10.1632/pmla.2020.135.1.188 Published online by Cambridge University Press

Our records are part of a longer history of the loss, gain, and reshaping of data that occur as bibliographic descriptions migrate from platform to platform and are converted from form to form. Along its path from acquisition to card catalog to OPAC, the novel both sheds and accumulates metadata. The University of Pennsylvania's card for *Pamela* has a detailed title that matches the title of Richardson's physical volume almost perfectly (fig. 1). The cataloger changed some of the punctuation and attempted to introduce a popular works title, "[Pamela. 1741]," before the actual title. There are other descriptive elements, including statements about edition, place of publication, publisher, and year of publication; the extent (two volumes); the book's length from the top to the bottom of the spine (seventeen centimeters); and its page size (duodecimo).

Like the paper card, electronic bibliographic records in the OPAC are updated over time; they are at once a finished product and a working draft. The OPAC record for *Pamela* contains more elements than the card, yet some data have vanished in the migration from card to electronic record (fig. 2).[12]

Our project attempts to include both the range of standardized fields in the digital record while also remembering the discursivity and the idiosyncratic range of knowledge of the paper catalog and including new

Staff View

| | | | |
|---|---|---|---|
| LEADER 02666cam a2200409 a 4500 | | | |
| 005 20181105155402.0 | | | |
| 008 850305m17411742enk 00000 eng d | | | |
| 001 992193473503681 | | | |
| 035 | | | **a**l (OCoLC)ocm16111904 |
| 035 | | | **a**l (OCoLC)16111904 |
| 035 | | | **a**l (CStRLIN)PAUG85-B8811 |
| 035 | | | **a**l (CaOTULAS)121252107 |
| 035 | | | **9**l ABE6311 |
| 035 | | | **a**l 219347 |
| 035 | | | **a**l (PU)219347-penndb-Voyager |
| 040 | | | **a**l PU **c**l PU |
| 090 | | | **a**l PR3664 **b**l .P3 1741 **i**l 03/06/85 CT |
| 100 | 1 | | **a**l Richardson, Samuel, **d**l 1689-1761. |
| 240 | 1 | 0 | **a**l Pamela **f**l 1741 |
| 245 | 1 | 0 | **a**l Pamela, or Virtue rewarded : **b**l in a series of familiar letters from a beautiful young damsel to her parents : now first published in order to cultivate the principles of virtue and religion in the youth of both sexes : a narrative which has its foundation in truth and nature, and at the same time that it agreeably entertains by a variety of curious and affecting incidents, is intirely divested of all those images which, in too many pieces calculated for amusement only, tend to inflame the minds they should instruct. |
| 260 | | | **a**l London : **b**l Printed for C. Rivington ... and J. Osborn ..., **c**l 1741-1742. |
| 300 | | | **a**l 4 v. ; **c**l 17 cm. (12mo) |
| 500 | | | **a**l Published anonymously. By Samuel Richardson. |
| 500 | | | **a**l Originally complete in 2 vols., the publication of a spurious sequel led him to continue it into 2 further vols.--Sale, W.M. Samuel Richardson, p. 11. |
| 500 | | | **a**l Vols. 3-4 have variant subtitle and imprint: In a series of letters from a beautiful young damsel to her parents, and afterwards in her exalted condition, between her and persons of figure and quality, upon the most important and entertaining subjects in genteel life ... publish'd in order to cultivate the principles of virtue and religion in the minds of the youth of both sexes / by the editor of the two first. London : Printed for S. Richardson and sold by C. Rivington ... and J. Osborn ... |

FIG. 2
The University of Pennsylvania Libraries' OPAC record for *Pamela*.

https://doi.org/10.1632/pmla.2020.135.1.188 Published online by Cambridge University Press

bibliographic data not present in either the paper or the online catalog. And END records also separate some of the data that the OPAC and the card catalog collapse. For example, eighteenth-century fiction often represents authorship in complex forms scattered across title page and prefatory material, but these representations collapse in the catalog, where known authors are almost always attached to records and linked to authority files. The END record for *Pamela* of course marks Richardson as the author of the novel, but it also captures the very different authorial and editorial claims made in the pages of the novel itself (fig. 3). In this way, the END record both represents *Pamela*'s own textual claims about its authorship and captures the familiar extratextual author claims that now adhere to it.

Because our primary dataset is made up of a set of bibliographic records, our dataset is by default organized by the individual works of fiction the records describe. But the idea of the work so shapes this form of bibliographic knowledge that it can make other forms of knowledge disappear, forms of knowledge that were not always confined to the boundaries of the work.[13] To suggest alternative ways of viewing the data that might show us how eighteenth-century books imagined themselves not just as finished works but as collections of elements that might be imaginatively unbound, we turn toward the paratextual element as the organizing principle for deriving new datasets.[14] In addition to our main, record-based dataset, we create tables of data in which rows are organized by aspects of novels rather than by the individual work, reorienting the data around the paratextual essays, authorial claims, epigraphs, marginalia, and other features found in the works of fiction in our dataset.[15] These datasets abandon the catalog record and the bibliographic work in favor of other forms. For example, our TSV dataset of paratextual essays creates a new row for each of the 1,448 paratextual elements—such as introductions, prefaces, and dedications—in our set of metadata; this dataset contains details about each paratextual element, including columns for the controlled-term type of paratext (whether preface, dedication, etc.) and a transcription

```
18626    <marc:datafield ind1=" " ind2=" " tag="599">
18627      <marc:subfield code="a">Author (paratext)</marc:subfield>
18628      <marc:subfield code="1">The editor</marc:subfield>
18629      <marc:subfield code="2">The editor.</marc:subfield>
18630      <marc:subfield code="3">Preface</marc:subfield>
18631      <marc:subfield code="5">Indeterminate</marc:subfield>
18632    </marc:datafield>
18633    <marc:datafield ind1=" " ind2=" " tag="599">
18634      <marc:subfield code="a">Author (paratext)</marc:subfield>
18635      <marc:subfield code="1">J.B.D.F.</marc:subfield>
18636      <marc:subfield code="2">Your most obedient, and faithful servant,
         • J.B.D.F.</marc:subfield>
18637      <marc:subfield code="3">Letter</marc:subfield>
18638      <marc:subfield code="5">Indeterminate</marc:subfield>
18639    </marc:datafield>
```

FIG. 3
This snippet of the END record for *Pamela* shows an example of encoding the authorship claims found in the novel's famous paratextual essays.

https://doi.org/10.1632/pmla.2020.135.1.188 Published online by Cambridge University Press

of the paratext's title if present, as well as information linking back to the work in which the paratext is found: its unique ID number, title, volume number, author, and publication date. Some works in our dataset contribute four or five items to this paratextual dataset, while some works contribute nothing.

Freed from the form of the work and the record, these transformed datasets imagine a textual environment in which prefaces, introductions, and dedications disregard the bindings of the novels to which they belong and speak across books to one another. Allowing paratextual essays to cross the boundaries of the books that hold them brings to view an eighteenth-century literary context in which literary criticism developed in prefaces as well as in the literary essays and reviews that appeared in periodicals. To give another example, our dataset on epigraphs offers information about the 719 epigraphs found among the works in our main dataset, linking transcriptions of the epigraphs and information about their origins and authors with standard information about the works of fiction in which they are found. The dataset can tell us how many books in it use William Cowper epigraphs and which works of Shakespeare are most popular for epigraphs.[16] The epigraph dataset temporarily detaches the epigraphs in our particular set of fictions from the books that hold them and transforms them into a corpus of their own. This transformation allows us to imagine what it would be like for a reader to temporarily experience the range of reference of the epigraphs detached from their contexts in works and volumes. Still, their tether to the original records allows users to connect changes in the epigraphs' range of reference to time, genre, and other bibliographic features.

A slow bibliographic metadata project, END relies on the time-consuming work of its many collaborators to create, check, remediate, and transform our relatively small dataset. Our process is inefficient from many points of view. But the nature of our work helps us explore some of the limits of the library catalog and the bibliographic description in order to imagine new possibilities for forms so familiar and well-worn that they and their accustomed uses can sometimes seem inevitable or invisible. Our project thus belongs to a centuries-long tradition created by catalogers, index builders, taxonomists, subject librarians, bibliographers, researchers, literary critics, and others whose work—much of it similarly slow—forms the infrastructure of humanistic scholarship.

## Notes

1. For example, the project Collections as Data has worked to create standard and best practices for making library data of all kinds available to multiple audiences and users (Padilla et al.).

2. MARC remains the primary form of bibliographic data in online public access catalogs.

3. MARC fields and subfields are standardized number fields for specific bibliographic information; 100, for example, designates the "personal name" (often an author) associated with a book, while 245 designates the "title statement." For a complete list of MARC fields, see "MARC 21."

4. The ideas, practices, and protocols of the project we describe here come from the collaborative work of our core researchers and contributors: Jeremy Culver, Sierra Eckert, Scott Enderle, Alexis van Eyken, Lynne Farrington, Mitch Fraas, Anne Garrison, Sam Herron, Nabil Kashyap, Anna Levine, Alice McGrath, Mayelin Perez, John Pollack, Charlotte Priddle, Beth Seltzer, Yumi Dineen Shiroma, Lindsay Van Tine, and Dan Traister, in addition to our summer undergraduate researchers, whose names and projects can be found at earlynovels.org or earlynovels.github.io.

5. Our primary work has been with the Singer-Mendenhall Collection of the English Novel held at the University of Pennsylvania Libraries' Kislak Center for Special Collections, Rare Books, and Manuscripts. We have worked with fiction in other repositories and collections as well, describing works of British and American fiction held at Swarthmore, Bryn Mawr, Haverford, the Library Company of Philadelphia, and the Fales Library at New York University in collaboration with librarians, faculty members, and student researchers at those institutions.

6. On paratexts, see Duncan and Smyth.

https://doi.org/10.1632/pmla.2020.135.1.188 Published online by Cambridge University Press

7. Despite their elaborate paratexts, eighteenth-century novels were first cataloged only by author and title—and sometimes by size and genre—in lists that early circulating libraries made of their holdings (Kaufman 14n30).

8. On the place of eighteenth-century fiction in information history, see Barchas; Valenza.

9. See the complete schema at github.com/earlynovels/end-dataset.

10. Drucker explains that capta with the characteristics of humanities documents would "have to embody ambiguity, complexity, fluidity, dynamic change, co-dependence, and other features of humanistic phenomena" ("Graphical Approaches" 248).

11. On reduction as a crucial element of work in literary studies, see Allison.

12. The record can be found at franklin.library.upenn.edu/catalog/FRANKLIN_992193473503681.

13. This is changing somewhat as the manipulation of digital records and the computational transformations of corpora increasingly organize bibliographic knowledge not just by the record and the work but by the volume and the page.

14. For an exploration of the myriad ways the book eludes the idea of the fixed and finished work, see Gillespie and Lynch.

15. As we struggled one summer with a solution to the problem of better representing the richness of our data for researchers who might not want to extract fields and subfields from the record sets themselves, Lindsay Van Tine came up with this simple but ingenious model of creating data subsets that centered on paratexts rather than on works.

16. For a visualization of the END epigraph data, see github.com/earlynovels/epigraph-visualization.

## Works Cited

Allison, Sarah. *Reductive Reading: A Syntax of Victorian Moralizing*. Johns Hopkins UP, 2018.

Barchas, Janine. *Graphic Design, Print Culture, and the Eighteenth-Century Novel*. Cambridge UP, 2000.

Drucker, Johanna. "Graphical Approaches to Digital Humanities." *A New Companion to Digital Humanities*, edited by Susan Schreibman et al., Wiley-Blackwell, 2016, pp. 238–50.

———. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly*, vol. 5, no. 1, 2011, www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Duncan, Dennis, and Adam Smyth, editors. *Book Parts*. Oxford UP, 2019.

Gillespie, Alexandra, and Deidre Lynch, editors. *The Unfinished Book*. Oxford UP, forthcoming.

Kaufman, Paul. "The Community Library: A Chapter in English Social History." *Transactions of the American Philosophical Society*, vol. 57, no. 7, 1967, pp. 3–67.

"MARC 21 Format for Bibliographic Data." *Library of Congress*, May 2019, www.loc.gov/marc/bibliographic.

Padilla, Thomas, et al. "The Santa Barbara Statement on Collections as Data, Version 2." *Collections as Data*, 2019, collectionsasdata.github.io/statement/.

Rawson, Katie, and Trevor Muñoz. "Against Cleaning." *Debates in the Digital Humanities 2019*, edited by Matthew K. Gold and Lauren F. Klein, U of Minnesota P, 2019, pp. 279–92.

Tsing, Anna Lowenhaupt. "On Nonscalability: The Living World Is Not Amenable to Precision-Nested Scales." *Common Knowledge*, vol. 18, no. 3, Fall 2012, pp. 505–24.

Valenza, Robin. "How Literature Becomes Knowledge: A Case Study." *ELH*, vol. 76, no. 1, 2009, pp. 215–45.

https://doi.org/10.1632/pmla.2020.135.1.188 Published online by Cambridge University Press