



## The Women's Print History Project

---

### The Ecology of Databases (feat. Lawrence Evalyn), *The WPHP Monthly Mercury*

Produced by Kate Moffatt and Kandice Sharren

Mixed and mastered by Alexander Kennard

Transcribed by Hanieh Ghaderi and Sara Penn

Music by Ignatius Sancho, "Sweetest Bard," *A Collection of New Songs* (1769), played by Kandice Sharren

Project Director: Michelle Levy (Simon Fraser University)

Moffatt, Kate, and Kandice Sharren, hosts. "The Ecology of Databases (feat. Lawrence Evalyn)." *The WPHP Monthly Mercury*, Season 2, Episode 6, 21 November 2021, <https://womensprinthistoryproject.com/blog/post/91>.

PDF Edited: 22 April 2024

---

This podcast draws on research supported by the Social Sciences and Humanities Research Council of Canada and the Digital Humanities Innovation Lab at Simon Fraser University.



Social Sciences and Humanities  
Research Council of Canada

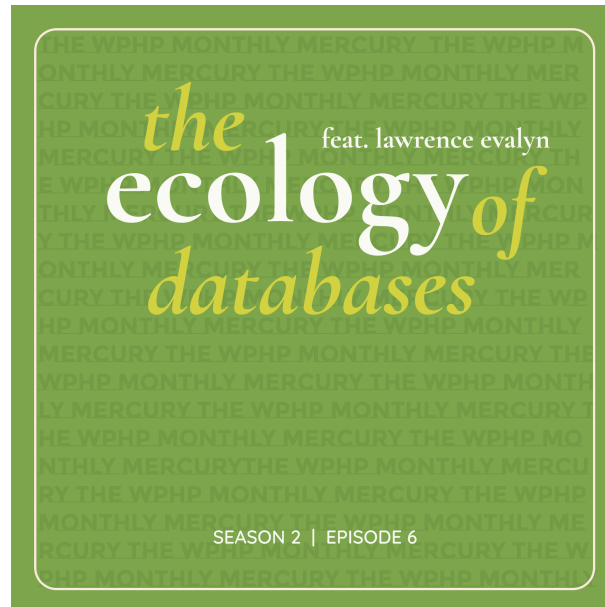
Conseil de recherches en  
sciences humaines du Canada

Canada



# The Ecology of Databases (feat. Lawrence Evalyn)

*Kate Moffatt and Kandice Sharren*



Why hasn't the third edition of Hannah More's *Coelebs in Search of a Wife* been digitized? Why doesn't *Google Books* group the different volumes of multi-volume works together in a single catalogue record? And, what do authors and pandas have in common? We bemoan the limitations of our various sources on a monthly basis, but this month we're digging into why they exist in the first place—especially why digitization can be so uneven.

In Episode 6 of Season 2 of *The WPHP Monthly Mercury*, "The Ecology of Databases," co-hosts Kate Moffatt and Kandice Sharren are joined by Lawrence Evalyn to learn more about the issue of uneven digitization. In addition to giving us the hard numbers about which titles appear in the ESTC, ECCO, *The Text Creation Partnership*, and *HathiTrust*, Lawrence puts forward his "charismatic megafauna" theory of authorship, shares moon prophecies and invitations to meetings about waterway management, and details the searching strategies he used during the WPHP Summer Readathon.

Lawrence Evalyn is currently a "pre-doc postdoc," both a Ph.D Candidate and a Teaching Postdoctoral Fellow in English at the University of Toronto, where he is affiliated with the Digital Humanities Network and the Data Sciences Institute. His dissertation, "Database Representations of English literature, 1789-99," measures and historicizes uneven digitization in four resources to examine how digital infrastructure shapes eighteenth-century studies, especially the study of women's writing. His collaborative digital humanities publications include "One Loveheart At A Time," an article on emoji in *Digital Humanities Quarterly*. He holds a Masters in English from the University of Victoria, where his M.A. essay, supervised by Robert Miles, looked for large-scale trends in late eighteenth century Gothic novels.

## WPHP Spotlights Referenced

“The Woman of Colour: Don't Break the (Attribution) Chain”

## WPHP Records Referenced

*The Woman of Colour* (title)

*Selections from the Letters, &c. of the late Miss Carter* (title)

*Tales Original and Translated from the Spanish. By a Lady.* (title)

## WPHP Sources Referenced

*Hathi Trust Digital Library*

*Eighteenth Century Collections Online*

*English Short Title Catalogue*

*Google Books*

## Works Cited

D'Ignazio, Catherine and Lauren F. Klein. *Data Feminism*. MIT Press, 2020.

Dominique, Lyndon J. “Introduction.” *The Woman of Colour*, Broadview Press, 2008, pp. 11–42.

Evalyn, Lawrence. “What Does Author Metadata Tell Us?: Counting English Women Writers in Four Eighteenth-Century Databases.” Unpublished.

Evalyn, Lawrence, C.E.M. Henderson, Julia King, Jessica Lockhart, Laura Mitchell, and Suzanne Conklin Akbari. “One Loveheart at a Time: The Language of Emoji and the Building of Affective Community in the Digital Medieval Studies Environment.” *Digital Humanities Quarterly*, vol. 14 no. 3, 2020.

Garside, Peter. “The English Novel in the Romantic Era: Consolidation and Dispersal”. *The English Novel, 1770–1829: A Bibliographical Survey of Prose Fiction Published in the British Isles*, edited by Peter Garside et al., vol. 2, Oxford UP, 2000.

Griffin, Robert J. “Fact, Fiction, and Anonymity: Reading Love and Madness: A Story Too True.” *Eighteenth Century Fiction*, vol. 16 no. 4, 2004, pp. 619–638.

Jockers, Matt. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

Levy, Michelle and Mark Perry. “Distantly Reading the Romantic Canon: Quantifying Gender in Current

Anthologies.” *Women’s Writing*, vol. 22 no. 2, Apr. 2015, pp. 132–155.

Moretti, Franco. *Distant Reading*. Verso, 2013.

Riddell, Allen and Troy J. Bassett. “What Library Digitization Leaves Out: Predicting the Availability of Digital Surrogates of English Novels.” *Portal*, vol. 21 no. 4, 2021, pp. 885–900.

### **Further Reading**

Bode, Katherine. *A World of Fiction: Digital Collections and the Future of Literary History*. U of Michigan P, 2018.

Cope, & Leitz, R. C. *Textual studies and the enlarged eighteenth century: precision as profusion* / edited by Kevin L. Cope and Robert C. Leitz, III. Bucknell University Press ; Rowman & Littlefield, 2012.

Gregg, Stephen H. *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge UP, 2021.


Harol, Corrinne, Brynn Lewis, and Subhash Lele. “Who Wrote It? The Woman of Colour and Adventures in Stylometry.” *Eighteenth-Century Fiction*, vol. 32 no. 2, 2020, 341–53.

Klein, Lauren. “Distant Reading After Moretti.” *Arcade: Literature, the Humanities, & the World*, 2018, [arcade.stanford.edu/blogs/distant-reading-after-moretti](https://arcade.stanford.edu/blogs/distant-reading-after-moretti).

Spedding, Patrick. “‘The New Machine’: Discovering the Limits of ECCO.” *Eighteenth-Century Studies*, vol. 44 no. 4, 2011, 437–53.

Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. U of Chicago P, 2019.


00:00:00 Lawrence Evalyn (guest) I increasingly think of the figure of the author as this kind of charismatic megafauna that when you come to environmental conservation, people always want to save the pandas and they don't want to save weird frogs [Kate and Kandice laugh]. And so you raise money for environmentalism by putting pandas on all of your flyers. And those are the charismatic megafauna that attract attention [Kate and Kandice laugh] to the problem. And I kind of feel like the figure of the author functions like this charismatic megafauna attract scholarly attention to a particular text.

00:00:42  [music playing]

00:00:51 Kandice Sharren (co-host) Hello and welcome to *The WPHP Monthly Mercury*, the podcast for *The Women's Print History Project*. The WPHP is a bibliographical database that collects information about women and book production during the eighteenth and nineteenth centuries. My name is Kandice Sharren—

00:01:06 Kate Moffatt (co-host) and I'm Kate Moffatt—

00:01:08 Kandice Sharren (co-host) and we are longtime editors of the WPHP and the hosts of this podcast. This season, we have some exciting, special guests to interview new research, to share and more stories to tell. Join us every third, Wednesday of the month, to learn more about the history of women's involvement in print.

00:01:26  [music playing]

00:01:34 Kate Moffatt (co-host) We have often talked about the limitations of the WPHP; the limits of available gender data, for example, or the scope of the project as a database that only captures titles from 1700 to 1836. But one of the biggest limitations and one that has come up time and time and time again [Kandice laughs] is our limited access to digitizations of the works that we're including in the database.

00:01:58 Kate Moffatt (co-host) As a quick refresher, we require either seeing a book in person or seeing a digitization of a book to consider it verified in our system. We don't explicitly capture data for whether a record has a digitization attached to it or not, but we do currently have 3,737 titles labelled as 'attempted verification.' This means a) we haven't seen a physical copy in a library, and b) there are no available digitizations of the work. So digitizations, to put it bluntly, play a pretty significant role in what we do, which is probably why we never shut up about it. [laughs]

- 00:02:32 Kandice Sharren (co-host) But limits can also be hidden: the unspoken or implied, or even unintentional biases of the people who create the resources we use are in some ways replicated in the WPHP. Their material and intellectual contexts limit them, which is reflected in what they record and make available. When we use their resources, we in turn are limited by what they contain. This is true for digital collections like *HathiTrust* and *Eighteenth Century Collections Online*, but also for the libraries and special collections that hold the physical copies needed to digitize them in the first place.
- 00:03:09 Kate Moffatt (co-host) Which brings us to what today's episode is all about. This month we are so excited to be joined by Lawrence Evalyn. Lawrence studies uneven digitization and its many causes, which means today's episode not only talks about the limits of the WPHP and its sources, but why some of those limits exist and how they came to be. Lawrence is currently a pre-doc post-doc both a PhD candidate and a teaching post-doctoral fellow in English at the University of Toronto, where he is affiliated with the Digital Humanities Network and the Data Sciences Institute.
- 00:03:42 Kate Moffatt (co-host) His dissertation, "Database Representations of English Literature 1789 –1799", measures and historicizes uneven digitization in four resources to examine how digital infrastructure shapes eighteenth century studies, especially the study of women's writing. His collaborative digital humanities publications include "One Love Heart at a Time", an article on emoji and *Digital Humanities Quarterly*. He holds a master's in English from the University of Victoria where his MA essay supervised by Robert Miles looked for large-scale trends in late eighteenth-century gothic novels.
- 00:04:16  [music playing]
- 00:04:25 Kandice Sharren (co-host) Thank you so much Lawrence for joining us and for sharing your in-progress work with us. We are so excited to talk to you about why so many of the limitations that we encounter in our sources exist and why they matter for projects like ours and yours. So, welcome.
- 00:04:46 Lawrence Evalyn (guest) Thank you so much. Yeah. I'm really excited to be here talking about it, especially because I feel like half of my project got done—I was imagining what would be the perfect bibliographic database and then I found *The Women's Print History Project* and it had everything! [laughs]
- 00:05:03 Kate Moffatt (co-host) Amazing!

- 00:05:06 Kandice Sharren (co-host) That's the most amazing thing to hear. [laughs]
- 00:05:10 Lawrence Evalyn (guest) Yeah. So since then, I'm just like, "I can't believe I went so long without knowing about this" [laughs]. I can't believe more things aren't like this. So I'm really pleased to be here.
- 00:05:19 Kate Moffatt (co-host) And that honestly makes so much sense to us. As we were thinking about this interview, we were like "digitizations they're so hugely important to the work that we do on the WPHP"—it's where we get a lot, if not most of our data, If we can't find a digitized copy of a book or a or hand check it in a library, then we can't verify it!
- 00:05:38 Kate Moffatt (co-host) And especially during COVID, but also as a team primarily based in Vancouver with limited travel opportunities, digitizations are how we do the majority of our work. We actually can't find things so often that we made a field for it called "attempted verification", which indicates that, despite our very best efforts, we could not verify the title against a digitization.
- 00:06:00 Kate Moffatt (co-host) So your research, which is focusing on the curation of digital resources and the gaps that we're struggling with, like Kandice mentioned, it's such an important question, but not one that would've necessarily occurred to us if we hadn't been dealing with these gaps in the digital archive on a basically a weekly basis [laughs]. So can you tell us about how you became interested in this topic? What kinds of gaps did you encounter?
- 00:06:26 Lawrence Evalyn (guest) Yeah, and I think in my case, I also found the gap first through trying to do something else and then became interested in how and why it was there. So, I first encountered "uneven digitization"—which is kind of the framework that I'm investigating is uneven digitization in mass digital archives—when I was actually trying to do a large-scale project on the female gothic.
- 00:06:53 Lawrence Evalyn (guest) I was really interested in whether all of those eighteenth-century stereotypes about the female gothic kind of actually held true for large bodies of women writers. And I was curious whether there was a legible male gothic, as well. And so the way I wanted to approach that was a classic digital humanities distant reading corpus-based project with a couple hundred gothic novels and key them to gender and just see if there were predictive features that could tell male and female gothic novels apart.

- 00:07:27 Lawrence Evalyn (guest) And I even managed to find good lists of what books I would want to look at, but it all fell apart at the corpus building stage in terms of getting clean digital texts. Anybody who's worked in the eighteenth century now sees the hilarious optimism of this quick—it was a one year master's program [all laugh]—and the plan was to get several hundred. [all laugh]
- 00:07:52 Kate Moffatt (co-host) Oh! Right. [laughs]
- 00:07:55 Lawrence Evalyn (guest) But I was reading all of this distant reading work that did have corpora of several hundred texts. And at the time I thought it was just a weird coincidence that they were so interested in twentieth-century or nineteenth-century texts. And I think I was still also in this more naive stage of encountering the "techno optimist" brand of DH, because that was a while ago as well. And so I hit that wall face first very hard. [all laugh]
- 00:08:24 Kandice Sharren (co-host) The earlier the better! [laughs]
- 00:08:26 Lawrence Evalyn (guest) Yeah, yeah. Early enough that I was able to do something else. I worked with bibliographic indexes actually as almost—
- 00:08:35 Kate Moffatt (co-host) Cool.
- 00:08:35 Lawrence Evalyn (guest) Yeah. Using bibliographic indexes basically as metadata about texts and seeing if that could still answer my questions. The other idea that's played a really big role in how my research has progressed from there is frankly work that increasingly historicizes those early techno optimist days of digital reading. And so especially there's work by Katherine Bode and Lauren Klein who do really interesting feminist-flavoured DH work that critiques especially how the Franco Moretti's and Matt Jonker's mode of digital humanities or distant reading work really perpetuates this ahistorical approach to literature,
- 00:09:22 Lawrence Evalyn (guest) and a huge part of how they kind of make their claims of having this direct and objective access to all of literary history is by concealing the corpus work, and concealing where their digital texts come from and even which texts they're looking at when they claim to look at the Victorian novel. And that made it really clear to me that actually there's this huge unexamined aspect of work on 'how do you build bodies of texts? What does it mean to be querying one set of digital



facsimiles or another?’ And that whole body of work honestly got more interesting to me than the original question about ‘was the female gothic, a thing that really happened.’

- 00:10:11 Kate Moffatt (co-host) Amazing.
- 00:10:12 Kandice Sharren (co-host) So in the research that you shared with us, you are kind of using the ESTC as your control, and you’re comparing that against various digitizations such as I think you have ECCO and *HathiTrust* and—
- 00:10:28 Lawrence Evalyn (guest) And the Text Creation Partnership. Yeah.
- 00:10:31 Kandice Sharren (co-host) So do you want to just very briefly summarize what you’ve noticed as the general trends in those?
- 00:10:42 Lawrence Evalyn (guest) Yeah. This is the stuff that I hope to eventually have an article coming out about because I think it’s incredibly difficult to know, just at this basic level, what books are in these different resources and what I found—I took this same sample, ‘same’ in definitely scare quotes there, of everything published in England between 1789 and 1799, in each of those four resources. And in theory, because it’s the same sample, they should have the same demographic breakdown if each of them were consistent or objective, whatever that would mean. But it would, at least it would be reassuring if they all agreed with each other.
- 00:11:35 Lawrence Evalyn (guest) So what I found was that first, the numerical differences in size were really striking. For that decade in the ESTC, there are 51,000 works published in England, 1789 to 99. But in ECCO there’s under 27,000, so slightly more than half of what’s in ESTC makes it into ECCO. The Text Creation Partnership, which is also a strict subset of ECCO, only has 525 for that sliver of time. And *HathiTrust* is in between the ECCO and the TCP. *HathiTrust* has a little more than 8,000. So part of what really got me interested once I realized those numbers are so different is that essentially selection has happened.
- 00:12:29 Lawrence Evalyn (guest) Some processes decided which half of the ESTC goes into ECCO. And some process decided which 1 percent of the ESTC gets hand-transcribed into the Text Creation Partnership, which is the only one that could be used for that distant reading corpus based kind of study is the Text Creation Partnership, which is *so* tiny as a sliver of everything produced during that decade. And my hypothesis

- going in was that given what I know about the world at large, that selection process probably would disadvantage women writers.
- 00:13:07 Lawrence Evalyn (guest) That's how I conceived of this as feminist influenced project was I thought I was going to go in and measure a historic mechanism of sexism. And I was really surprised to find not that. So the hard part of the research was that I did go through and manually look at every author name for all 52,000 things [all laugh] and write down whether I could tell it was a man or woman because—
- 00:13:42 Kate Moffatt (co-host) People on the podcast cannot see me right now, but I just made a brain exploding motion with my hands. [laughs]
- 00:13:49 Lawrence Evalyn (guest) I sometimes forget about this as a major step in the research process. It honestly wasn't as time consuming as you would expect because each first name I was able to do just once through. So there's something like 5,000 by John. And so all of those 5,000 are easy to identify—
- 00:14:13 Kate Moffatt (co-host) Out they go! [laughs]
- 00:14:13 Lawrence Evalyn (guest) John, I recognize that's a male name. And to be precise there, the thing that I'm evaluating is an implied gender based on the author attribution as recorded in the database. Which from just looking at a whole bunch of these, a) it's usually actually quite possible to identify a probable gender from the information there. And b) the name in the database entry very rarely matches the name as it appeared on the title page of the work itself—
- 00:14:49 Kate Moffatt (co-host) Oh!—
- 00:14:49 Lawrence Evalyn (guest) that it really seems like in most cases, the database is filling in a blank of some kind and telling us the name of a specific person for works that didn't have a full name on the title page itself.
- 00:15:06 Kate Moffatt (co-host) I hope I'm not jumping ahead here, but I find that really fascinating because in the WPHP [Kandice laughs], we have a contributor field where we'll put who the actual authors are, but then we have a signed author field where we will indicate how things are written on the title page.
- 00:15:21 Lawrence Evalyn (guest) Yeah. That is jumping ahead but I also find it so exciting—

that was one of the things that I found so *frustrating* working with these other resources is that it is so hard—

00:15:34 Kate Moffatt (co-host) They don't do that [laughs]—

00:15:36 Lawrence Evalyn (guest) I wish more places would record that. And I think there's this way where once somebody has done an authorship attribution, that name kind of overwrites in the database the author information as provided by the title page. And it kind of varies the original printed attribution. And I understand why there's not a lot of point in finding out who wrote things if their names don't go into databases so that people know what they wrote. Really, you could have two fields—

00:16:10 Kate Moffatt (co-host) Right. Yeah! [laughs]—

00:16:12 Lawrence Evalyn (guest) as the WPHP does. That's the solution: have two fields.

00:16:18 Kandice Sharren (co-host) Separate that information out!

00:16:20 Lawrence Evalyn (guest) Yeah. I'm about to describe to you an increase in female authorship in the database, but I can't concretely at this stage know whether that has to do with works that are unattributed in one database, like pseudonymous in some way, becoming attributed in later databases.

00:16:45 Lawrence Evalyn (guest) For the most part, it looks like the attribution match between ESTC, ECCO, and TCP, which makes sense, because it's all kind of one ecosystem. But it's definitely an open question for me. How much of the phenomena I observe are due to selection pressures of which texts are in these databases versus how much is due to the same texts being recorded differently database to database?

00:17:15 Kate Moffatt (co-host) Okay.

00:17:17 Kandice Sharren (co-host) That is a good question.

- 00:17:18 Lawrence Evalyn (guest) Which I find a fascinating question especially because, okay here's my kicker. The real good stuff is I found in the ESTC 1789 to 99, the percent authored by women, 3 percent, which is so low.
- 00:17:37 Kate Moffatt (co-host) That's not very high. [laughs]
- 00:17:39 Lawrence Evalyn (guest) It's not very high. So 3 percent in the ESTC. In ECCO 4 percent. *HathiTrust* 5 percent. TCP 22 percent.
- 00:17:52 Kate Moffatt (co-host) That's a big jump.
- 00:17:53 Kandice Sharren (co-host) Yeah. [laughs]
- 00:17:55 Lawrence Evalyn (guest) Yeah. Which is almost perfectly the reverse of what I expected. I expected to see more than 3 percent female as the baseline, because our numbers for women in the novel specifically are above 20 percent for this decade. And I was like, "I don't know how much of it is novels. How much did they write other things?"
- 00:18:19 Lawrence Evalyn (guest) You have to remember, I did start this project as a person who only the year prior was like, "yeah, I'll just get 300 clean texts of gothic novels."
- 00:18:27 Lawrence Evalyn (guest) You learn a lot about print culture, print history by trying to do projects and learning what parts are hard.
- 00:18:37 Kandice Sharren (co-host) Yeah. So just to contextualize those percentages a bit, before you joined the call and before we started the interview, we did a quick search on the WPHP just for how many titles we had between 1789 and 1799. So those same dates with women listed as authors in some way, whether we know who those women are or not.
- 00:19:01 Kandice Sharren (co-host) And our kind of raw number was 2,104. And then we have of that set, 1,689 have been verified. So that means we've either seen a digitization or we've seen a hard copy somewhere. So how many in the ESTC do you have—that 3 percent did you say?
- 00:19:24 Lawrence Evalyn (guest) Yeah, I'm doing that quick math—so 3 percent of what I have in the ESTC would be 1,532.

- 00:19:34 Kandice Sharren (co-host) Interesting.
- 00:19:35 Kate Moffatt (co-host) Interesting.
- 00:19:37 Kandice Sharren (co-host) We have a little bit more over 500 more titles than what you have.
- 00:19:43 Lawrence Evalyn (guest) And that could be because I do have from the ESTC, I do have 2 percent of titles that have some information on the title page, but I wasn't able to infer gender. And 21 percent are unsigned, nothing in the author field. And so from that 23 total percent that I didn't assign at gender to maybe some of those are where your extras come from.
- 00:20:09 Kandice Sharren (co-host) Yeah. So I wonder if that's also just differences in how we determine or search for authorship or list authorship. But that's an interesting discrepancy. Okay. I was curious about how that would line up.
- 00:20:23 Kate Moffatt (co-host) And actually, okay, sorry. I use the ESTC a lot, and I won't ask you questions about this right now because I think that would be really mean it's not stuff that you work on. So I won't do that. But I use the ESTC a ton for finding books that were published, printed, or sold by women printers, publishers, and booksellers, because they let you search the publisher field, which is so amazing!
- 00:20:45 Kate Moffatt (co-host) So when we actually search not for just how many books do we have that have the women as authors in that decade, or those 11 years, when we search for a total, how many do we have? We have another 600 books on top of that that also have women were involved in the production somehow. So either as editors, compilers, or printers, book sellers, whatever.
- 00:21:09 Kandice Sharren (co-host) And that's a number that's going to increase too because you're still working on it.
- 00:21:12 Kate Moffatt (co-host) Oh it changes constantly. I'm still working on that. Yeah, absolutely. But I find it so interesting—that's the other thing too. That gender gets captured in different fields in different ways too.
- 00:21:25 Lawrence Evalyn (guest) Yeah. And that was one of the decisions I had to make actually was deciding how I wanted to define gender and how I wanted to treat texts where the gender wasn't necessarily marked in the author field. But when you looked at the whole

title, for example, it would refer to the author with a gendered pronoun. And so, if you were picking up the book, it was telling you right from the start something about the gender of the author.

- 00:21:57 Lawrence Evalyn (guest) And I ended up going with, if I could, from the database record of my spreadsheet to see those kinds of markers, imagining myself as a scholar interested in studying a work and thinking about the gender of its author I decided to take that information into account. And so that mostly actually happened to books that turned out to be by men, is—
- 00:22:21 Kate Moffatt (co-host) Oh interesting—
- 00:22:21 Lawrence Evalyn (guest) that it would be ambiguous in some way, but then the title itself would include a male pronoun. And so I would say, okay, so this is a male author, which might just be because so many things are by men [Kandice laughs]. So one of the interesting things that goes hand in hand with that increase in female authorship, which actually mirrors something that Michelle Levy and Mark Perry saw in their paper on the *Norton* anthology is that there's more and more women as the databases grow smaller and more selective, but there are also more and more men.
- 00:22:58 Lawrence Evalyn (guest) That it goes from 49 percent male to 65 percent male to 72 percent male to 76 percent male. So it's just a little steady climb. It's not as dramatic as the increase in female authorship, but it is definitely the case that an increase in female authors is *not* a decrease in male authors. And it's for actually the same reason that they found that they found that *Norton* anthologies just got longer and that's how they could give more pages to women and also give more pages to men. Here it's just that—
- 00:23:37 Kate Moffatt (co-host) I'm sorry, that's so funny to me. [all laugh]
- 00:23:41 Lawrence Evalyn (guest) It just feels so yeah, it's such a non-solution to the problem.
- 00:23:46 Kate Moffatt (co-host) It's a non-step forward almost. It isn't but it's just funny to me anyway. Sorry, please go on. [laughs]
- 00:23:55 Lawrence Evalyn (guest) Yeah, no. I found it really funny when I encountered it and I found it kind of helpful to see something not quite parallel happening here, but basically the

works that are disappearing are works that I think of under the broad umbrella of ‘authorless.’ I increasingly think of the figure of the author as this kind of charismatic megafauna that when you come to environmental conservation, people always want to save the pandas and they don’t want to save weird frogs. [Kate and Kandice laugh]

- 00:24:49 Lawrence Evalyn (guest) And I kind of feel like the figure of the author functions like this charismatic megafauna attract scholarly attention to a particular text. And so there are these works that are different in their publication contexts that some of them have totally blank title pages, some of them are by organizations that they’re corporate authored in some way. Some of them are clearly bulletins, advertisements, laws, maps.
- 00:25:19 Lawrence Evalyn (guest) I found a strange number of things in the ESTC that are just invitations. That the title is just something like, “sir, your presence is requested at...” [Kate and Kandice laugh]. And there’s details of some place and time they’re supposed to show up. And they were clearly just handed to a specific individual. There’s a list of these twenty people are all expected to be at this meeting and they’ve printed it for convenience in some way. And then it’s made its way into the ESTC and into ECCO. Many of those are in ECCO as well actually, is how I know—
- 00:25:51 Kate Moffatt (co-host) That’s crazy to me—
- 00:25:52 Lawrence Evalyn (guest) what they look like. But it’s not a work with an author. And so—
- 00:25:58 Kandice Sharren (co-host) How would you assign an author to that? [laughs]
- 00:26:01 Lawrence Evalyn (guest) Yeah. And they’re really weird meetings too. They’re consulting on proposals to change, to dredge the Thames or there are a lot of things about waterway management [Kate and Kandice laugh]. So, all of these works, what they share is this lack of a figure of an author. That you have to find your way into them through an interest in the text itself some other way.
- 00:26:32 Kate Moffatt (co-host) In waterways, you have to be interested in waterways. [laughs]

- 00:26:35 Lawrence Evalyn (guest) Yeah. You have to be interested in waterway management or ineffective cures for scurvy. I've got some moon prophecies [Kate laughs]. See, and this is where I think actually all of this is really cool. And so I'm increasingly interested in what is valuable and interesting about these texts, even though they don't have this central biographical figure of the author.
- 00:27:07 Lawrence Evalyn (guest) And so, because at first I think it feels like, is it even a problem that we have unevenly digitized these kinds of non literary or sub literary materials? And I think there is a problem to it. It might not feel as urgent as a problem of if this was reinforcing the social marginalization of women, author lists isn't a social category, it's only a category of texts. And so there's less urgency to processes that marginalize these kinds of works.
- 00:27:45 Lawrence Evalyn (guest) But I still think that there's a loss involved or that it's worth sometimes going and taking a look those things and keeping them in mind as we think about what constitutes normal or common in the period as we're trying to contextualize other works that do have their charismatic panda authors.
- 00:28:07 Kandice Sharren (co-host) So what would the percentage of authorless texts be that you've identified in the ESTC?
- 00:28:16 Lawrence Evalyn (guest) So in the ESTC it's 47 percent authorless, which is part of why I think this is a category of text worth paying attention to because it's almost half of everything.
- 00:28:27 Kate Moffatt (co-host) Yeah. That's a lot.
- 00:28:28 Kandice Sharren (co-host) That's a lot! [laughs]
- 00:28:32 Lawrence Evalyn (guest) Yeah. And that drops to only 2 percent in the TCP, the smallest one. And it steadily declines across the four that each one has less and less of these broadly authorless works, as the resources get smaller and more selective. And so that's why I argue that this is a selection force that authorless of a text is predictive of its failure to be invested in, in a subsequent round of digital infrastructure.



- 00:29:07 Kandice Sharren (co-host) So let's talk a little bit about what gets lost when we don't make those authorless texts available. You mentioned the very well known anonymous title, *The Woman of Colour* when we were kind of having our email conversation leading up to this chat and that might be a great place to start, but thinking more broadly, what do we not see when we're not digitizing things that don't have authors or clearly identifiable authors?
- 00:29:39 Lawrence Evalyn (guest) Yeah. I think *The Woman of Colour* is a really good entry point to this because a lot of the things that are authorless do look like this, "sir, your attendance is requested at a waterway meeting" [Kate and Kandice laugh] but that's sort of on one end of the extreme and quite a lot of these authorless works are also exactly the kinds of things that we do pay a lot of attention to and see a lot of value in, in traditional literary scholarship, like *The Woman of Colour: A Tale* which is a really great and teachable text as well.
- 00:30:16 Lawrence Evalyn (guest) And I think something we've talked about a little bit, or something that we talked about wanting to talk about is how a lot of the scholarship on *The Woman of Colour* is really focused on authorship, is really focused on who the author is. Lyndon Dominique makes a really interesting case in the introduction to that *Broadview* edition that it could have been written by a woman of colour.
- 00:30:44 Lawrence Evalyn (guest) And I think one of the things I like about his take is that he's kind of more interested in what's the value of taking that possibility seriously than he is in actually answering the question of authorship in itself. And I think there is real value in, it is helpful to remember that when you don't know anything about an author, that means you don't know anything about an author.
- 00:31:07 Lawrence Evalyn (guest) And so they might not have been a white man. I think this is something more broadly that I think we really want, especially when it comes to grappling with social marginalization in literature, I think we really want there to be clear causal links between the social identity of the author as a person, the content of the work as a text, and also the social and political commitments of the reader.
- 00:31:39 Lawrence Evalyn (guest) We want all three of those things to go nicely together. And I think that they don't [laughs]. And I think especially that when we don't know the identity of the author, that doesn't mean we have to let go of exploring the implications of the content of the work and exploring the political impacts on the reader or readership of that book.

- 00:32:08 Kandice Sharren (co-host) Yeah. Well, and I think *The Woman of Colour* is interesting too, because we don't know a specific historical individual who wrote it, but if we look at the title page information— there is author information there, author information that Peter Garside, I believe, has suggested is perhaps false or not completely accurate.
- 00:32:33 Kandice Sharren (co-host) But it is linking us and pointing us towards three other titles that we could go look at and it's claiming an authorial relationship with those titles. So even if we don't necessarily have that historical individual that we can be like, “this is who they were and what their identity was and how they fit into society,” we do have information about how this text is positioning itself in relation to other texts.
- 00:33:02 Lawrence Evalyn (guest) Yeah, I think that idea that even if you don't know who the author is, you're thinking about the author as a person. And the book is trying to relate with to your idea of the author. Robert Griffin has a lot of really interesting work on that, that I've been finding more and more useful—especially, he has this article on *Love and Madness, a Story Too True* [Kandice laughs] and as you're reading it, you are constantly in this position of trying to evaluate whether it is a true story or not, and who wrote it. And that's a great core part of the reading experience.
- 00:33:42 Kate Moffatt (co-host) Cool. I love it! [laughs]
- 00:33:44 Lawrence Evalyn (guest) Yeah. I got halfway through that title and sort of forgot how fun it is. *Love and Madness, a Story Too True*. [all laugh]
- 00:33:52 Kate Moffatt (co-host) I was going to say, you sound like me because I'm obsessed with exclamation points in eighteenth-century novel titles. I think they're just the best.
- 00:33:58 Lawrence Evalyn (guest) Oh, they're so good.
- 00:33:59 Kate Moffatt (co-host) I'm obsessed with *The Three Monks!!!* exclamation point exclamation point exclamation point. [laughs]
- 00:34:05 Lawrence Evalyn (guest) Oh, I love that one!
- 00:34:05 Kate Moffatt (co-host) Anyways, we're getting mildly off track.

- 00:34:07 Lawrence Evalyn (guest) We've definitely gotten off topic, but I do think this is an example of looking through books for something other than the author, right? That the other one of the other things you can use to get interested in a book is its title. And, I think especially when you want to think about the gothic novel and the way that it operated through these really cliché and formulaic titles and subtitles, I feel like part of that is because they aren't author driven, they're driven by something else. And so it manifests in this other way. And so I think these are some of the kinds of questions that can get overshadowed by still strongly biographical approaches to literary history.
- 00:34:54 Kandice Sharren (co-host) Lawrence, a lot of your work as we've been discussing treats digital and physical collections themselves as historical artefacts, which means thinking about the cultural and material contexts of their production and how those contexts prevent these collections from being representative of book production more generally. So I just want to take a couple minutes to follow this line of thinking and consider the conditions that shape digital collections a little bit more explicitly.
- 00:35:22 Kandice Sharren (co-host) So for example, *Gale's* digitizations are based on microfiche from the 80's, which is why they're such a nightmare to read [laughs] and TCP is limited by the amount of labour that goes into transcription. What are some of, for you, the most striking ways that the contexts of digital archives have shaped what they contain? And do you have any thoughts about how to make curation more equitable or if not more equitable, more intentional?
- 00:35:53 Lawrence Evalyn (guest) This is a great question. And it's also going to be a tricky one because a lot of this is thinking that I'm still in the middle of, and I think two frameworks come to mind that I want to take a little time to think through. And the first one is this idea of nations and nationalism, and the other is commerce and commercialism. And so thinking about how nationalism plays out in these specific concrete histories of these resources, I can't get out of my head the way that the ESTC defines 'English' literature, what is 'English' enough that they're laying claim to it, it's their job to collect.
- 00:36:44 Lawrence Evalyn (guest) And it's everything printed in the British Isles or territories governed by Britain, including the United States all the way up to 1800, in all languages. Or, anything printed anywhere, at least partly in English or British vernacular. And so that has this really striking melding of nation and language as defining the boundaries of

- ‘Englishness’ that feels like it’s really trying to claim and intentionally conflate English and British in a way that makes sense for the British library to want to do,
- 00:37:26 Lawrence Evalyn (guest) especially when I found that the creation of the ESTC chronologically really lines up with and seems to be a reaction to some really exciting cutting edge work at America’s Library of Congress, where they were developing MARC as an encoding standard for how to record bibliographic information systematically. And the ESTC really explicitly does not use MARC [Kate and Kandice laugh]. It uses its own different thing.
- 00:38:04 Kandice Sharren (co-host) [laughs]. That’s really interesting actually, because one of the things that we ran up into because the ESTC gave us a huge chunk of their data that we’ve imported. And then we had to go through and filter it out and clean it. And one of the things we came across a lot of were titles in French that listed London as the place of publication, but it was a false imprint and it was in fact, a French translation of a work that was listing the original publishers on the title page. And those are all also included in the ESTC. So it’s interesting that that is also being folded into this. It’s not just things in English. [laughs]
- 00:38:49 Lawrence Evalyn (guest) Yeah. They do say also on their list that they include false imprints, that claim to fit one of their criteria—
- 00:38:58 Kate Moffatt (co-host) So interesting—
- 00:38:59 Lawrence Evalyn (guest) and with the, yeah, with the false imprints, I can kind of see it as half a pragmatic decision of it is really hard to know what books are lying to you because most of them are lying to you about something, it seems. [Kate laughs]
- 00:39:12 Kate Moffatt (co-host) Books are liars. Publishers are liars. This is something we talk about constantly on the podcast.
- 00:39:18 Kandice Sharren (co-host) Everyone’s a liar. We’re liars too. [all laugh]
- 00:39:20 Lawrence Evalyn (guest) Yeah. I feel like every spreadsheet I have has this asterisk of “this is what’s in the database subject to lies by the book makers.” So something that I’m still sort of chewing on, but that I find enduringly interesting and worth thinking more about is the way that this expansive definition of Englishness kind of is still

- naturalized or feels acceptable. That it's this way of claiming, for example, anything published in India would be English under this framework.
- 00:39:57 Kate Moffatt (co-host) Which is interesting. That has interesting implications, I guess, in terms of what they're claiming as British or English, there's 290 results.
- 00:40:08 Lawrence Evalyn (guest) For things printed in Calcutta. Yeah. And so I think that's especially interesting because of the way that also some of the time in the eighteenth century, England was trying to pretend that India was not governed by Britain. It was just subject to this commercial relationship with the East India Company. Yeah, that's interesting the way that the ESTC retroactively is acknowledging that these were lies—
- 00:40:42 Kandice Sharren (co-host) Yeah—
- 00:42:43 Lawrence Evalyn (guest) that this is a distinction without a difference between England and the East India Trading Company in that period. But while also still naturalizing the claim in itself. That's something I'm still thinking about and finding compelling with the role of nation states in shaping what texts get siloed together and therefore can be worked on together versus what texts don't have an easily accessible institutional home. And so it's like a 'damned if you do damned if you don't' scenario as well. That if it's excluded from the institutional resources, it's harder to study, but if it's included, it's making this claim to it that feels uncomfortable.
- 00:41:42 Kate Moffatt (co-host) There's almost a colonization energy to it. That you have to kind of try and untangle.
- 00:41:48 Lawrence Evalyn (guest) Yeah. Which I think is most stark in the ESTC's claim that American literature is just English literature all the way up till 1800 [laughs]. That they're like, we just don't recognize that anything changed in 1776. It all goes in.
- 00:42:06 Kandice Sharren (co-host) Okay. So that was one of the contexts you wanted to talk about. What was the other one?
- 00:42:10 Lawrence Evalyn (guest) So that was the way that you see how these archives are shaped by the role of nations and the other is this idea of commerce. It's hard to pick just one or two examples of how money has shaped what is and isn't available to be consulted as a digital facsimile one way or another. And they're all extra interesting because

people want to talk about them even less than they want to talk about the role of nationalism in the creation of these resources.

- 00:42:45 Kate Moffatt (co-host) Interesting!
- 00:42:46 Lawrence Evalyn (guest) By going back to that period of techno-optimism from early DH, there's two projects that have had actually a really big impact on how digital techs are used: *Google Books* and *Project Gutenberg*. And they're really characterized by this *wild* techno optimism that *Project Gutenberg* announced that they wanted to grow their collection to 1 million free e-books and distribute them to 1 billion people for a total of one quadrillion e-books given away by the end of 2015. That was their goal. They have not hit 1 million e-books. They have about 66,000, which is still a lot.
- 00:43:33 Kandice Sharren (co-host) That's quite a lot.
- 00:43:34 Lawrence Evalyn (guest) And *Google Books* used to routinely say that their plan was to scan every book in the world. Then they went through and removed all those claims from their website. [all laugh]
- 00:43:50 Kate Moffatt (co-host) Never mind, we changed our minds! [laughs]
- 00:43:53 Lawrence Evalyn (guest) Yeah. And *Google Books* I think is especially worth looking at because Google did all of this scanning for *Google Books*, because they thought they could make money off of it. There's these really interesting interviews where they talk about how the way Google search outranks its competitors is by knowing what information people find more valuable. And man, the way that they talk about books is *wild*. They say things like, "the best information is trapped in books." And "it's useless unless you can find a way to extract and schematicize it." And so—
- 00:44:41 Kate Moffatt (co-host) Wow. That's such a Google point of view! [all laugh]
- 00:44:45 Lawrence Evalyn (guest) Yeah. So *Google Books* is this mass digitization project that begins basically because they want a really large corpus of text. And they think that it will enhance their search engine. And the way that they start doing the digitization, which blatantly

violated copyright, was by forming partnerships with libraries where they would literally drive semi trucks up to libraries and fill them with books and then drive them to centralized scanning centers where humans would turn the pages of all of the books.

- 00:45:18 Kandice Sharren (co-host) Those hands that you get in the scans. [laughs]
- 00:45:21 Lawrence Evalyn (guest) Yeah. You see hands, you see all these other human traces. But so Google does all of the scanning because they think that it will make them money and then they get sued because they have basically built for themselves this monopoly. Because originally they also wanted to sell out-of-print e-books, or orphan works, that basically works that nobody seemed to know who owned them and therefore you can't get them.
- 00:45:58 Lawrence Evalyn (guest) Google was just going to sell you those e-books at some negligible price. And in some ways it was a kind of neat solution to this problem of orphan works. But on the other hand, and the reason they got sued is it would give them a perpetual monopoly on orphan works that no one could compete with except by illegally scanning millions of books [all laugh]. An incredibly expensive process that no one could do again.
- 00:46:28 Lawrence Evalyn (guest) So this is how Google ends up with all of these page images and they don't get out of it quite what they wanted. They don't get this free money forever from orphan works, but they do get *Google Books*, which does exist. And presumably it helped their search engine the way that they wanted. And they sort of quietly wound down the scanning. They still scan some books, but not at these incredible truckloads at a time rate.
- 00:46:52 Lawrence Evalyn (guest) And they definitely don't talk anymore about how they're going to scan every book in the world. And the way that this really impacts scholars is that *HathiTrust* is those pictures Google took, but now for us [Kate and Kandice laugh]. That when a library agreed to let Google photograph their book in exchange they got the photographs of their own books. And so they immediately encountered this kind of network problem.
- 00:47:24 Lawrence Evalyn (guest) This is, I don't know how well I can explain this, but it's a similar problem you see with social network websites or social media websites. That the main thing that's good about Facebook, if anything is good about Facebook, is that a lot of

people are already there. And it's hard for a new social media website to compete when it doesn't already have the people there. And so—

- 00:47:53 Kate Moffatt (co-host) You need the numbers.
- 00:47:55 Lawrence Evalyn (guest) Yeah, that you need to get over this critical mass numbers to increase the usefulness of the thing that you've made. And that analogously goes with mass digitization that a resource is more useful the more comprehensive it is. And so when you had a lot of different libraries that each had their own images of just their books that Google wanted to photograph from them that has some usefulness, but it's not a place that you can just turn to immediately expecting to find a book that you want to read. But when all of them get pooled together, it becomes genuinely a second *Google Books*.
- 00:48:34 Kate Moffatt (co-host) Right. And one that's much easier to search. [laughs]
- 00:48:38 Lawrence Evalyn (guest) Yes. Oh my gosh. You can really tell.
- 00:48:41 Kandice Sharren (co-host) The metadata is much better. Yeah.
- 00:48:43 Lawrence Evalyn (guest) Yeah. You could really tell when you try to use *Google Books*, they do not care about books.
- 00:48:49 Kate Moffatt (co-host) Trying to find particular editions on *Google Books*.
- 00:48:52 Kandice Sharren (co-host) Volume one is here, volume four is here. Where's volume two and three, who knows?
- 00:48:57 Lawrence Evalyn (guest) Oh my gosh.
- 00:48:58 Kate Moffatt (co-host) They're starting to make changes. They're starting to connect you. There's, "here's other additions of this book," "here's other volumes of this book." But sometimes they're completely different years of publication and that's not what I'm looking for.



- 00:49:10 Lawrence Evalyn (guest) But to try to return to this idea of commercial impacts on these kinds of resources, what interests me about *Google Books* as the secret source of *HathiTrust* is that I feel like it gets at this problem that doing this work is expensive. And it often feels like we're trying to find ways to build incredible things without having to spend money on them.
- 00:49:44 Lawrence Evalyn (guest) And once I start thinking about that, this idea that you're trying to do this very expensive thing, and also you have to make money off of it, I start to feel more uncomfortable with the way that ECCO is run by a for-profit company. And it's so complicated as a scholar who absolutely relies on ECCO for everything I do to want to point out, it's a little gross, the way that ECCO markets itself for its comprehensiveness and its completeness.
- 00:50:16 Lawrence Evalyn (guest) And somehow manages to do that while selling ECCO and ECCO 2 as separate products, which, if you look in their websites have the identical descriptions. That when you just want to know how many things are in ECCO they give you the biggest number they can possibly justify and they put that number for both ECCO and ECCO 2, even though for them to be purchasable separately, they must contain different amounts of stuff [laughs]. Right?
- 00:50:45 Lawrence Evalyn (guest) And there's this sense where one of the reasons it's hard to know things like how many works are by women in the eighteenth century is because there's commercial incentives to make the contents of these resources more opaque. And there are commercial incentives to overemphasize their completeness and their size.
- 00:51:14 Kate Moffatt (co-host) And I think this ties in so strongly to things we've talked about before, but it's this idea that people really—I say people, students, researchers; depends on how familiar you are with it—that there's this false narrative of “you can find everything online. If it's not in that database, it doesn't exist.” These are things that we've been kind of trained to believe that if it's a database, a) it's correct and b) it holds everything that it could possibly hold. You know what I mean? And it's just so—
- 00:51:45 Kandice Sharren (co-host) Or it holds everything that's important.

- 00:51:48 Lawrence Evalyn (guest) Yeah. I think it's *especially* that bit. I think that we associate this maybe with a student-y perspective of, oh, undergraduates, I've read this in multiple reviews of ECCO. There's a lot of published work that say variations on undergraduates can be seduced into thinking that if it's not in ECCO, it doesn't exist. But I think even as scholars, we can be seduced into thinking if it's not in ECCO, it's not important. That it's not worth the hassle of trying to dig it up some other way. It will be fine if we just use ECCO.
- 00:52:21 Kandice Sharren (co-host) Yeah. No, or alternatively, you could think it's important, but if you can't access it, you can't do anything with it. So it ends up kind of getting swept under the rug or ignored.
- 00:52:33 Kate Moffatt (co-host) And we wonder all the time, we're like, "oh, I can't fight a digitization of this." Who's not prioritizing it? Is it just, is it low on the librarian's list of things to digitize? Is it lost in a box somewhere that nobody accesses? It becomes this question that we have, because we're like, "no, we're trying to include these and prioritize them or treat them equally, at least with the things we can equally access." It's like "who is not prioritizing this? Is this an issue?" [laughs]. Why is this not digitized compared to something else? Right.
- 00:53:09 Lawrence Evalyn (guest) Yeah. And I think that's always my question: Why is this not digitized compared to something else? And what I find always rewarding—
- 00:53:16 Kate Moffatt (co-host) This is your whole thing! This is what you do!
- 00:53:18 Lawrence Evalyn (guest) Yeah. And what I find really rewarding and I think we're well prepared to grapple with the answers that we find. But I think we're really well trained. I don't know, I say "we" [laughs]. Especially in the scholarly circles that I move in, which is really influenced by feminist scholarship and decolonial scholarship and antiracist scholarship. People are putting a lot of work into trying to identify and resist certain mechanisms of deprioritization.
- 00:53:53 Lawrence Evalyn (guest) But I'm also finding a lot of things that are much stranger than that. So a really interesting article by Alan Riddell and Troy Bassett looked at novels from 1838 using, I think it was a Garside list actually. They took this list of what they considered every novel published in that year and looked at which ones of them were available in digital facsimiles. So one of the biggest influential factors that

they found actually was that the British library didn't digitize any multivolume novels. And so—

- 00:54:33 Kandice Sharren (co-host) Wow!
- 00:54:34 Kate Moffatt (co-host) Interesting! Because it was more labour?
- 00:54:36 Kandice Sharren (co-host) But the three volume novel was the thing!
- 00:54:39 Kate Moffatt (co-host) The thing!
- 00:54:44 Lawrence Evalyn (guest) They don't even speculate as to why it could be—
- 00:54:46 Kate Moffatt (co-host) That's so shocking.
- 00:54:48 Kandice Sharren (co-host) They didn't get around to it. [laughs]
- 00:54:51 Lawrence Evalyn (guest) Yeah. They were in a different part of the building. They have a different kind of bibliographic code and that section of their code didn't get lined up to be scanned or whatever. It's both clearly a gender blind omission, right? They didn't get scanned because they were three volume novels, which probably has something to do with three volume works, their three volumeness, but—
- 00:55:26 Kandice Sharren (co-host) [laugh]. Well, and there are more work to scan probably. They take more time. Let's do the ones that are quick and easy first.
- 00:55:34 Kate Moffatt (co-host) But think—*Pride and Prejudice* has three volumes. What are the chances they didn't digitize [laughs] the three volumes of *Pride and Prejudice*?
- 00:55:45 Kandice Sharren (co-host) There is no complete digitization of the 1813 edition of *Pride and Prejudice* out there. Fun fact. Yeah. Nope. But these are just like 1838, right? So this—

- 00:55:58 Lawrence Evalyn (guest) Just 1836. I had misremembered. They looked specifically at the year, 1836. And so my guess is that there might be really big omissions. I mean, other organizations might have scanned them. It might be that for 1836 *HathiTrust* has them all. But it's still striking how we can see how that could really easily impact women disproportionately while also not really being sexism.
- 00:56:34 Kate Moffatt (co-host) Not appearing to be gendered. [laughs]
- 00:56:37 Lawrence Evalyn (guest) Yeah. Although—
- 00:56:40 Kandice Sharren (co-host) But if women are writing more novels than anything else and a whole bunch of novels aren't digitized, then suddenly a whole bunch of work by women has not been digitized.
- 00:56:50 Kate Moffatt (co-host) It's just not as explicit
- 00:56:51 Kandice Sharren (co-host) Disproportionately. Yeah.
- 00:56:53 Lawrence Evalyn (guest) And that does loop back around to where it's like, okay and so why do we think novels aren't as important to collect into libraries in the first place? How do other things that seem to be just about the relative worth of different kinds of works can kind of be like, "oh, well we don't think novels are as important because they're associated with women."
- 00:57:23 Lawrence Evalyn (guest) And so it seemed like it was just about novels, but it circles back around. I think that's especially the case in later time periods well, past the eighteenth century with things like children's fiction. It's underrepresented in *HathiTrust* because it's under collected by academic libraries because it's for children and college students aren't children, but then are we happy with that as the outcome?
- 00:57:56 Kandice Sharren (co-host) We've been really grappling in the last few months with the fact that there is not a ton of children's literature digitized because we're trying to work through all the titles that we haven't tried to verify yet. And most of it is children's literature because that's the stuff where like someone tried to tackle it and was like, "oh, this is a headache." And then gave up because there are no digitizations and it's a pain.

- 00:58:23 Kate Moffatt (co-host) And it's interesting because we've talked about this a bit on the podcast as well; children's lit, specific genres, do have more scholarship about them than others. And I would say children's lit is one of those, novels is another, so it's poetry, but that doesn't necessarily carry over to digitizations [laughs] of what you have actually like available for ascertaining that data or making sure that it's correct. So we're stuck working with just print bibliographies from scholars who were really focused on a particular genre. So that's such an interesting, I don't know, disconnect, I guess.
- 00:59:03 Lawrence Evalyn (guest) Yeah. And what you say there, I find really interesting also from this idea that I think it's easy to feel like research infrastructure exists for researchers. So if we're interested in children's literature, which I think people increasingly really are, therefore our infrastructure such as our digital archives should be built with those needs—
- 00:59:27 Kate Moffatt (co-host) Support that [laughs]—
- 00:59:27 Lawrence Evalyn (guest) in mind. But actually these archives come to us downstream of other forces, many of which don't have to do with scholarly priorities and have to do with completely separate things. This is where I'd be curious to look more into journals in the realms of library science and archival studies.
- 00:59:56 Lawrence Evalyn (guest) I'd be curious if actually it's linked to historic moment of academic libraries separating themselves from public libraries. That would make sense to me as, "oh if you want children's literature go to the public library, this is an academic library and we have more serious kinds of texts." And my impression is that Google went disproportionately with academic libraries because they were the ones who were interested in having scans of their books as the sort of repayment.
- 01:00:26 Kate Moffatt (co-host) So interesting.
- 01:00:27 Kandice Sharren (co-host) Well that's interesting too, because the Osborne Collection, which was the source for a huge amount of our children's literature data, is held in the Toronto Public Library, not one of the University of Toronto libraries.
- 01:00:40 Lawrence Evalyn (guest) I was, that's a—

- 01:00:41 Kandice Sharren (co-host) That's a really interesting connection. [laughs]
- 01:00:43 Kate Moffatt (co-host) Yep. That's a good point.
- 01:00:46 Lawrence Evalyn (guest) Yeah. I had the Osborne Collection specifically in mind when I was thinking that even when children's literature has been collected, it seems like its home has been more naturally in public library systems. Because I think the New York Public Library also has some really excellent resources for children's literature.
- 01:01:05 Kate Moffatt (co-host) So before we have to be done this wonderful conversation that we've been having; you participated in our first readathon this summer, which was really fun and really cool. And you posted about some delightful and strange finds in the WPHP. And we were just really curious about your searching strategies.
- 01:01:27 Kate Moffatt (co-host) What were you looking for? How did you find it? Which fields did you really like to search in or use? What similarities or differences did you notice between your process of random sampling and searching the WPHP. Be our test subject [laughs]. Tell us about using the WPHP because we're so so curious.
- 01:01:53 Lawrence Evalyn (guest) I'm really glad you asked this because it was such a pleasure to use and really interesting also to use. I used it in two different ways. So first I wanted to read something— basically I didn't want to start with a totally random find because sometimes those can be a challenge to get through.
- 01:02:13 Kandice Sharren (co-host) Yes, they can. [all laugh]
- 01:02:15 Lawrence Evalyn (guest) So I wanted to start with something that I knew would be enjoyable and not too long. I always wish I could search by number of pages or filter by length in some way. Nobody implements this. And it's because it would be really challenging. Even when you have a collocation for a book it's not easily convertible to a number of pages.
- 01:02:39 Lawrence Evalyn (guest) So I understand why that field isn't in these databases and it's probably not worth putting in. But I was able with the WPHP to filter for one volume works, which put a useful cap on the length of what I was going to find. And I've also been increasingly interested in reading letter collections, just because I've been reading some and they're strange and interesting.

- 01:03:08 Lawrence Evalyn (guest) So I typed in the word ‘letters’, filtered to one volume and filtered to female authors. And I think I just picked the first one that was under 300 pages. I think I just scrolled through, opened up the individual entries and the first one that wasn’t unbearably long, I read. Which was, a really intriguing posthumous volume memorializing a young woman who had died, edited by her fiancé.
- 01:03:38 Lawrence Evalyn (guest) And there were two things that I found really neat about it: one was that it didn’t have any full letters, it just had little extracts—
- 01:03:51 Kate Moffatt (co-host) Oh!—
- 01:03:52 Lawrence Evalyn (guest) of nice turns of phrase. And he mentioned that he was intentionally cutting out all of her jokes and her friends would think it was weird how serious she sounds in this book, but he wanted to remember her this way.
- 01:04:04 Kate Moffatt (co-host) Oh! [laughs]. He pulled out all of her jokes?
- 01:04:09 Lawrence Evalyn (guest) Yeah, it made me really want to know what kinds of jokes she made in her letters.
- 01:04:15 Kate Moffatt (co-host): I know!
- 01:04:13 Lawrence Evalyn (guest) But a lot of her—
- 01:04:15 Kandice Sharren (co-host) No jokes that were fit for print. [all laugh]
- 01:04:25 Lawrence Evalyn (guest) Yeah! But the other thing that I really liked is that the preface that he wrote also really emphasizes that at the end of this volume of her letters are a bunch of his poems and he makes all these apologies like they were written in a frenzy of grief and my excessive feeling, and my friends told me I shouldn’t do it [Kate and Kandice laugh]. And so I tried not to, I’ve taken most of them out. I took out the worst stuff, but I couldn’t help myself.

- 01:04:54 Lawrence Evalyn (guest) This depth of pain has to be expressed through poetry in this way. Which was really also quite touching. Except then the poems are so boring [Kate and Kandice laugh]. It was this really strange—this is what I love about reading things coming in a little bit blind is I went through this whole volume, really becoming convinced of the depth and sincerity of his grief, especially because he selects these letters for places when she's offering consolation to other people, who've experienced a death in the family.
- 01:05:09 Lawrence Evalyn (guest) And so it's clear that he's reflecting on what she would say to him about her own death, if she was still alive. He's really in it. And then he has these poems that are just this insipid heroic couplets—
- 01:05:50 Kate Moffatt (co-host) Oh no!—
- 01:50:51 Lawrence Evalyn (guest) for pages and pages with all of these allegorical virtues and it feels so cookie cutter—
- 01:06:02 Kate Moffatt (co-host) And I'm dying because your fiance died and your publishing, some of her letters and your move is to publish really bad poetry? Is to take advantage of this? “But you know what, I've always wanted to be a published poet. I'm going to stick a bunch of my poetry at the end of my dead fiance's letter collection?” [all laugh]
- 01:06:21 Kandice Sharren (co-host) So that people will buy it!
- 01:06:25 Lawrence Evalyn (guest) Well, all of his friends told him not to!
- 01:06:26 Kate Moffatt (co-host) This is why his friends said don't do it! [laughs]
- 01:06:32 Lawrence Evalyn (guest) Yeah. But I also feel like it helped me think through a little bit, what does the sincere poem look like? Because I do buy—
- 01:06:42 Kate Moffatt (co-host) Okay—
- 01:06:43 Lawrence Evalyn (guest) that he was sad when he wrote these poems. I'm sure there's also this self-aggrandizement aspect of wanting to publish them anyway [Kate and



- Kandice laugh]. But he does hold them until the end. And he does tell you that you don't have to read them. [Kate and Kandice laugh]
- 01:06:57 Kate Moffatt (co-host) I'm against my better judgement— I'm actually appreciating this man. [laughs]
- 01:07:06 Lawrence Evalyn (guest) Yeah. Right. We can, I totally see where you're coming from in a kind of gross move. But what I found really fascinating was actually thinking through, again, this idea of almost thought experiments of literature. I found it really interesting to just look at the poem and entertain the possibility that it was an expression of sincere feeling.
- 01:07:34 Lawrence Evalyn (guest) A lot of his rhetoric kind of matched things that Wordsworth later says in the preface to *Lyrical Ballads* of poetry as this spontaneous overflow of feeling. And it's like, what does it mean to look at heroic couplets of allegorical virtues and to take them seriously as a spontaneous overflow of sincere feeling and so I didn't quite get there—
- 01:07:58 Kate Moffatt (co-host) And sorrow. And grief—
- 01:07:58 Lawrence Evalyn (guest) but I enjoyed the challenge [all laugh]. So that was my first search. And as I said, I feel like I had a really positive experience with the book itself because in part I was able to filter for something I wanted to learn more about something readably short, that was letters and had a specifically female author, wasn't just printed by a woman publisher or bookseller.
- 01:08:24 Lawrence Evalyn (guest) For the second thing I read, I wanted to do something more like some of my other work where I have been just using random number generators to get sets of texts and then trying to read them together and think about, okay, if all I knew about the 1790s was a scurvy cure, a prophecy about the moon, Erasmus Darwin's *Zoonomia*, and, *The Vicar of Wakefield*, what do I know about the 1790s? If these are the texts that I have to interpret it with? So I really enjoy random sampling as this way to explore literature separately from that classic author as charismatic megafauna mode.
- 01:09:11 Lawrence Evalyn (guest) But it's often really hard to take a random sample because of all of the forces that cause these resources to conceal their scope and contents. The WPHP was really easy because they have sequential unique identifiers. So I was able to just find what's the biggest number that anything has been given as a unique ID, use a

random number generator for something that's between one and that number. And I got something that was a valid book in the collection and it only took ten seconds.

- 01:09:46 Kate Moffatt (co-host) Amazing!
- 01:09:46 Kandice Sharren (co-host) I'm so glad that us making the title IDs searchable and visible to people has paid off because we had to fight for that.
- 01:09:58 Lawrence Evalyn (guest) Really?
- 01:09:59 Kate Moffatt (co-host) We did.
- 01:09:59 Kandice Sharren (co-host) Yeah. It would be really useful to be able to search by title ID. And our developer was like, "but why would you possibly need to do that?" [laughs]. And we had to come up with imaginary scenarios. So I'm really glad that it was useful for you. Thank you. [laughs]
- 01:10:15 Kate Moffatt (co-host) Our developer's great about that, by the way, he'll always ask us like, "okay, why do you want it?" And we're like, "okay, we have to actually explain this to you." And it makes us think through it in a different way, which is cool. So, anyway. Go on.
- 01:10:27 Kandice Sharren (co-host) But anyway, I'm extra glad that we fought for that now.
- 01:10:31 Lawrence Evalyn (guest) Yeah. Especially because I so often want to be able to recommend to people, "you should random sample too." And then I try to think about how, and it's like, "well, I only have my list of this ten years of ESTC records because I asked nicely for it from the ESTC directly." And then I had to go and assign them my own unique identifiers to get the sequential numbers.
- 01:10:54 Lawrence Evalyn (guest) And the spreadsheet even for the 10-year period is too large for my computer to open, so I actually have to use secondary software. It's like, "I can't in good faith suggest that other people do this." But with the WPHP anybody could really straightforwardly take a random sample. And the work I found was really cool. It was a collection of tales from later than I usually study, it was the eighteen-teens.

- 01:11:25 Lawrence Evalyn (guest) And so some of what was fun for me was like, “oh right, there’s more than just 11 years of fiction” [Kandice laughs]. And it had a lot of interesting gothic tropes, but being repurposed in this more classically Romance way. But so I really enjoyed that book of tales. It kind of felt like the random number generator gods were smiling down upon me [laughs] or maybe women just write better books because I’ve definitely received much worse things from processes like this before.
- 01:12:03 Lawrence Evalyn (guest) But I think the broader link from both of those two search experiences actually does touch on this idea that these interfaces are negotiations with entities like developers and that when you’re building a system, you’re also building the experiences someone can have and the questions they can answer with that system. And I do have sympathy for the developer that I’m really aware that every additional field you add—
- 01:12:31 Kandice Sharren (co-host) Absolutely—
- 01:12:32 Lawrence Evalyn (guest) is expensive and it increases the fragility of the system and the complexity. That you do have to make decisions and not include everything that you could possibly want include. It’s just really striking how much I can see when I’m using the WPHP in contrast to, I think—like I don’t want to be mean to the genuinely excellent resources that are the ESTC and ECCO. So using the WPHP in contrast to something like *Google Books*, you can tell so clearly in these infrastructural ways that it’s built for this kind of precision and historicism that *Google Books* just has no interest in.
- 01:13:23 Kate Moffatt (co-host) Well, I like that. [laughs]
- 01:13:26 Kandice Sharren (co-host) [laughs]. We’ll take it. Thank you.
- 01:13:27 Kate Moffatt (co-host) Is it really self-serving to say I like that? [laughs]
- 01:13:31 Lawrence Evalyn (guest) That’s fair. I guess I don’t know, maybe, no, I’m sure. Actually, you guys must think about this in I’m really curious about it. So the WPHP, though, isn’t one of the objects of inquiry of my research because I made the decision not to really look at the women-only resources that do exist.

- 01:13:56 Lawrence Evalyn (guest) That there's a lot of these projects, which I think for purposes other than trying to evaluate the overall demographics of historical print, it's clearly really useful to increase the accessibility of these works. There's a lot of other people's research questions that they can really help with, but it does feel like there's something strange or uncomfortable about the fact that knowing how many works are published by women, for example, as a number in and of itself can't help us understand their place and broader print culture—
- 01:14:35 Kate Moffatt (co-host) Oh, we talk about this all the time—
- 01:14:37 Kandice Sharren (co-host) [laughs]. Yeah.
- 01:14:38 Kate Moffatt (co-host) How we can't say women published this percentage of books in a particular time, because we don't have—
- 01:14:48 Kandice Sharren (co-host) Which is why when I knew you had that percentages, I was like, "oh, we need to compare now. I need to know what, how we stack up." [laughs]
- 01:14:55 Kate Moffatt (co-host) With all of these obscure women that we're including in the WPHP. There are also obscure men who haven't been recovered very well, whose data is also hard to find, but we can't speak to that because we focus on women specifically. So it's so easy for me to sit in my angry little bubble [laughs]. Why are women so hard to find? But sometimes it's worth considering that obscure men can also be hard to find.
- 01:15:25 Kandice Sharren (co-host) And that we do have a lot of information about some specific high profile men who ran a major book selling business in London or something, but that guy who like had a print shop in some small town in like Lancaster, we don't know any information about him either.
- 01:15:42 Kate Moffatt (co-host) Right. And that it makes it hard for us to look at things and be like, given that we probably don't know about all these other obscure men who really didn't amount to much who don't have anything to show for them or documentation or whatever [Kandice laughs]. Given that we don't have any of that—

- 01:15:56 Kandice Sharren (co-host) [laughs]. A mood.
- 01:15:59 Kate Moffatt (co-host) [laughs]. Will I ever amount much? But this quantitative data we're creating can't really be compared to say like the same kind of data for men because we don't have it. Right. So anyways, yes. One of the, one of the limitations of the WPHP.
- 01:16:16 Lawrence Evalyn (guest) Which I think though is a limitation of research or history itself, right? That for certain kinds of questions, you have to know so much in order to be able to contextualize them, that I was able to get a percentage for a tiny 11-year sliver in a really narrow geographic range. I'm just England, not even Great Britain. I don't do Scotland, which would be a whole different world of print, right? It's incredibly, incredibly narrow. And the only way I was able to get these few numbers to put in context with each other was an unreasonable amount of manual grinding.
- 01:17:02 Kandice Sharren (co-host) Yeah, you have to set limitations. If we were trying to do, if we've got 3 or 4 percent of all titles from the period we're looking at to kind of extrapolate from your data and we're trying to do men too—I'm already unsure if we will ever be done, but we would never, ever, ever be done with the resources that we had. But I think it's also just interesting to know kind of that hard number of well, "we have this many titles in the database, which is this many times more than what we thought we would have starting out."
- 01:17:38 Lawrence Evalyn (guest) I can't call up the details right now, but the ESTC actually made a really similar misestimation of their scope that they had this plan of, they would finish going through everything, I want to say in 10 years [laughs]. It did *not* take 10 years. And I think that they had underestimated the scope of what they were doing by an order of 10. It was something really large.
- 01:18:08 Kate Moffatt (co-host) Wow.
- 01:18:09 Lawrence Evalyn (guest) And it was actually one of the things that gridlocked some of the earlier attempts to get it funding and get it going. Is that a lot of your questions about: how feasible is the project? How much data can you collect? How much data is it feasible to record about each thing? Those questions are really influenced by the number of titles that you're going to have to do it for. But the reason you want to do the ESTC is so that you can know the number of titles that there are.

01:18:40 Kate Moffatt (co-host) Totally. Yeah. Right.

01:18:41 Kandice Sharren (co-host) One of the core questions. [laughs]

01:18:46 Lawrence Evalyn (guest) And so, and they were also wildly wrong about it. I can't find the details, but I'm really certain that they were extremely wrong.

01:18:54 Kate Moffatt (co-host) Yeah. And I do think that also speaks actually to the importance of the WPHP as a database as well. It's literally correcting these ideas that Michelle works on women's book history and she thought we'd only have 2000 or 3000 titles, that says something right. That she, that she even had that estimate in the first place, right? It says a lot about the narrative that we kind of had going into the project.

01:19:14 Lawrence Evalyn (guest) Absolutely.

01:19:15 Kate Moffatt (co-host) And we've now corrected that and it's an ongoing process.



01:19:18 Kandice Sharren (co-host) Yeah. Well, and it makes sense if all you know are the handful of canonical titles that we read—why would you think there are that many books during this period? So you would never have encountered them.

01:19:32 Kate Moffatt (co-host) Yeah. If you had never experienced, if you'd never experienced or encountered a cookbook that went into 50 editions before, why would you think that multiple of those exist? [laughs]. Which they do, FYI. It's so much fun chatting with you about everything.

01:19:48 Lawrence Evalyn (guest) Thank you so much. I have really enjoyed it.

01:19:50  [music playing]

01:19:58 Kate Moffatt (co-host) Our conversation with Lawrence drove home the fact that the majority of digitization initiatives are shaped by unseen factors that aren't necessarily beholden to scholarship or research interests. Why did the British Library only digitize one volume novels from 1836 when any self-respecting nineteenth-centuryist would tell you that the three volume novel is widely understood as the dominant form? We are, as Lawrence put it, downstream of

- that work and those initiatives, which means sometimes we're working with what already exists in digitizations, because that is what is available. And that can have ongoing implications.
- 01:20:32 Kandice Sharren (co-host) By amalgamating data from various resources, about as many books as we can find evidence of whether or not we find digitizations of them, the WPHP aims to offer a fuller picture of the print landscape during the long eighteenth century. For example, our choice to include anonymous works that indicate female authorship means that some of the large body of unattributed works is represented in our data.
- 01:20:58 Kandice Sharren (co-host) And, while we do aim to include all titles produced by women run firms, which means that some titles by the corporate authors that Lawrence talked about will be included, the importance of authorship to many of our sources means that it has shaped which titles we have identified and been able to include. We are also aware that our own contribution will be limited by the gendered and geographical scope of our project. Our focus on women's involvement in print limits our ability to understand their contributions in a wider context, even as it corrects cultural narratives about who engaged with print and how.
- 01:21:37  [music playing]
- 01:21:48 Kate Moffatt (co-host) This has been the sixth episode of Season 2 of *The WPHP Monthly Mercury*. If you're interested in learning more about what we discussed today, we've compiled a list of suggestions for further reading and links to some relevant entries in the WPHP, in a blog post that you can find at [womensprinthistoryproject.com](http://womensprinthistoryproject.com). You can also find us @theWPHP on Twitter and on Instagram @womensprinthistoryproject.
- 01:22:12  [music playing]
- 01:22:22 Lawrence Evalyn (guest) [outtakes, part 1] But I think sort of the broader link from those two different searching experiences have left my brain. [all laugh]
- 01:22:33 Kate Moffatt (co-host) [outtakes, part 2] I don't think I've ever said 'digitizations' that many times in my entire life.
- 01:22:41 Kandice Sharren (co-host) It's also a really hard word to say: digi-ti-za-tion.

01:22:43 Kate Moffatt Digitizations, digitizations. By the sixth time I was saying it, I was like...  
(co-host) 'digitization!' [Kandice laughs]

01:22:55 Kate Moffatt Without sounding like I'm thinking about it that hard.  
(co-host)

01:22:58 Kandice Sharren You know, just trying to be like, "yeah, I can say digitization. No problem."  
(co-host)

01:23:03 Lawrence Evalyn [outtakes, part 3] Oh, did I say 'a' or did I say 'one'?  
(guest)

01:23:05 Kate Moffatt You said 'a.' [laughs]  
(co-host)

01:23:11 Kate Moffatt [outtakes, part 4] I guess it's complicated. Right? You can't compare or consider  
(co-host) this qualitative data we're creating. How qualitative can it be if—

01:23:19 Kandice Sharren Quantitative.  
(co-host)

01:23:20 Kate Moffatt You can't compare it to—oh, dang. Yes. Do I mess that up every time?  
(co-host)

01:23:26 Kandice Sharren Often. [laughs]  
(co-host)

01:23:27 Kate Moffatt There's another blooper for us. [laughs]  
(co-host)