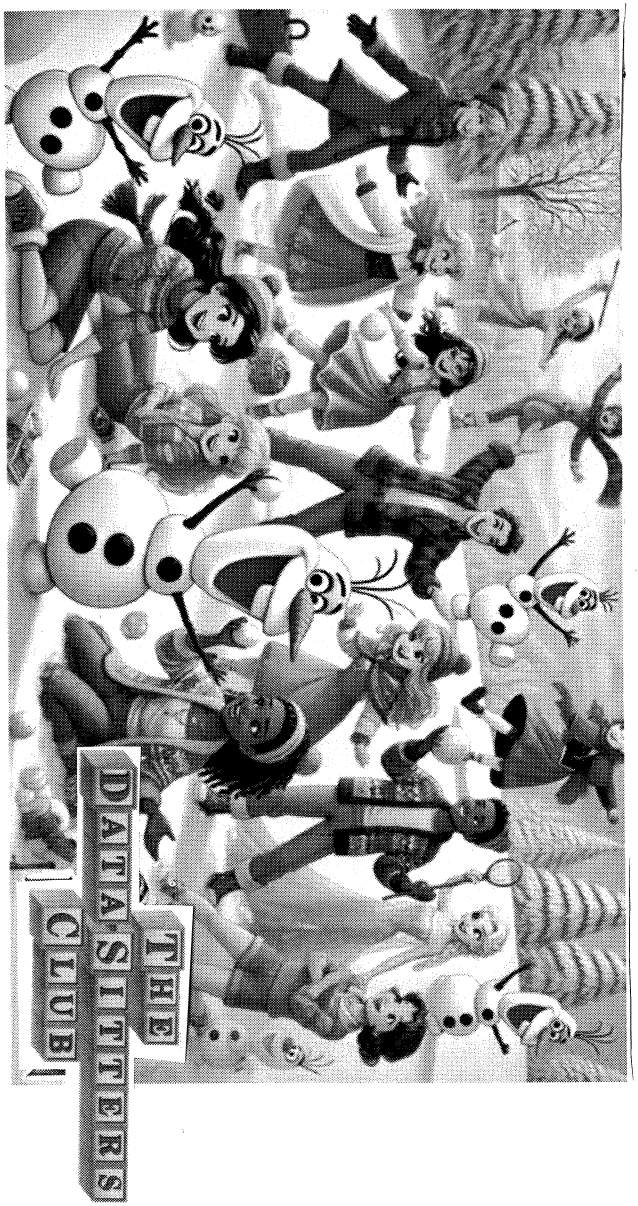
https://datasittersclub.github.io/site/

'Zine #1

THE DATA SITTERS CLUB



So You Want To Build A Corpus?

by Lee Skallerup Bessette, Katherine Bowers, Maria Sachiko Cecire, Quinn Dombrowski, Anouk Lang, and Roopika Risam

Who we are

First, we should tell you about The Data-Sitters Club, how it works, and who we are. It all started one day when Quinn Dombrowski was on vacation in Las Vegas and started getting nostalgic about Ann M. Martin's iconic series about girlhood in the upper-middle-class American suburbs of the 1990s. There's been very little scholarship written on the series, and Quinn started wondering what you could get from applying digital humanities computational text analysis tools and methods to that corpus. And it'd be even better to share those findings alongside a detailed description of how you go about the process of doing the text analysis, spelling out all the steps and decisions along the way! Quinn found some friends on DH (digital humanities) Twitter who wanted to join, and together, we started The Data-Sitters Club.

(I mean, you were building this whole time! HAHA!)

Gee, thanks ...

... Now what?

That's what the rest of the DSC books are for!

Not sure where to start?

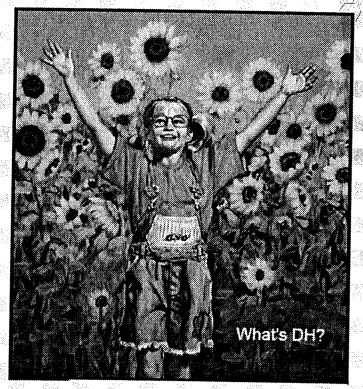
https://datasittersclub.github.io/site/

Learn more!

bitly

DATA-SITTERS

little tl;dr #1



DH Curious?

Roopika Risam, Katherine Bowers, Maria Sachiko Cecire, and Anouk Lang

bitly

https://datasittersclub.github.io/tldr/books/tldr1/

It. Is. Time. To. Start.
BUILDING!!!

https://datasittersclub.github.io/site/dsc2.html

Quinn's OCR Tips

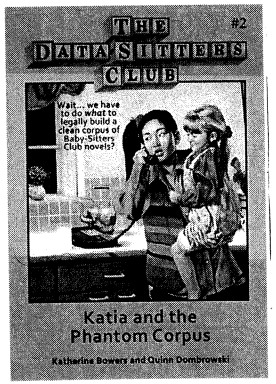
If you need to do some OCR yourself, here's some tips from someone who's spent a lot of time doing it:

Actually scan your sources. With a scanner. At minimum 300 dpi (dots per inch – it's a measure of how much detail the image captures). More like 400-600 dpi if your books include a really small font, like for footnotes. Yes, the technology is improving, and sometimes you can get better-than-total-garbage with photos from your phone, as long as they're not skewed (taken at an angle), the lighting is good, etc., but still, most phone pictures are still only 72 dpi, and it's hard to position your phone directly above a book and not cast a shadow. Just use a scanner.

Scan your sources in grayscale, especially if you're going to be using ABBYY FineReader. While all the OCR algorithms actually use on binarized images (black & white – where everything in the image is either black or white, according to some threshold you or the software defines), you can go from grayscale to B&W, but not the other way around. Even though the OCR algorithm itself involves a binarized image, the algorithms used for layout analysis (i.e. figuring out where the text is on the page, whether it's one column or two, whether there's tables or running headers or page numbers, etc.) are more nuanced. Also, both ABBYY FineReader and the open-source Tesseract software include pre-processing steps before they perform the OCR, including binarizing images using a sensible threshold that cuts down on the noise in the image. For instance, if you run a B&W scan of a two-page spread through Tesseract (i.e. an image where the binarization has happened at the time when you did the scanning), you'll end up with some gibberish from when the OCR algorithm tried to "read" the shadow caused by the binding.

Save your scans as .tif files, which are uncompressed and don't lose any of the data in the image to make the file size smaller. A 300 dpi grayscale scan of a two-page spread of a Baby-Sitters Club book (like you'd get when scanning it, assuming you don't want to just cut the book's spine and run it through a sheet-feeder scanner – which is viscerally disturbing, but a much more time-efficient option) is close to 7 MB, and there's around 80 such scanned images per book, which works out to around half a gigabyte per book for the page images. If all you want is good OCR, don't feel like you need to keep all these image files for the long-term: you're not responsible for library-quality preservation for the books. Once you're satisfied with the OCR quality, you can let them go and delete them.

Use ABBYY FineReader (on Windows). Ideologically, we all (especially in digital humanities) support open-source software. Various digital humanities projects have focused on improving OCR quality for Tesseract (e.g. Laura Mandell et al's Early Modern OCR Project which trained an earlier version of Tesseract for early modern typefaces, and the Open Islamicate Texts Initiative Arabic-script OCR Catalyst Project, which is providing a more user-friendly workflow for Arabic and Persian OCR based on Tesseract). But the fact is, while character-by-character recognition (at least for English) is basically identical between FineReader and Tesseract, FineReader does a lot of "common-sense" things that make your life easier:



(This is much better!)

Want to "do" DH? First, you need data, or a corpus, to do DH to!

However...

You Must Follow the LAW!

What exactly is copyright?

Most people, if asked casually, would probably say they know what copyright is. But then there are moments when cracks begin to show. If you read fanfic, for example, you'll likely come across disclaimers like "no copyright infringement intended" – almost like some legalistic twist on "no offense intended".

Here's the deal: anyone who makes almost any type of creative work immediately has a set of *exclusive* rights pertaining to that work.

The "creative" part is important: you can't copyright facts. Fan wikis (like the one we scraped for metadata about the BSC books in DSC Mysterious Mystery #3: Quinn and Lee Clean Up Ghost Girl Data Humbug) depend on this: there's nothing copyrightable about the fact that Ann M. Martin wrote Kristy's Great Idea in 1986, it was narrated by a teenage girl named Kristy Thomas, and she refers to her friends Claudia Kishi, Mary Anne Spier, and Stacey McGill. So all that information, plus a plot synopsis, the ISBN number, publication date, and various other details can be included on fan wiki pages. But the fan wiki can't post the full text of the book, because *that's* the creative work *itself*.

Similarly, a simple list of ingredients and preparation instructions isn't copyrightable, but the further the recipe skews towards personal narrative, the more likely it is to be eligible for copyright. But even then, copyright would protect only the narrative parts (the original expression) and people would still be free to copy the fact and instruction parts.

You can't just do whatever you want!

What's fair use?

The UK (where my fellow Data-Sitter Anouk Lang lives) and former British colonies (e.g. Canada, Australia, New Zealand, India, Singapore) have as part of their copyright law something called "fair dealing" – which enumerates a specific and finite list of uses of in-copyright works that **do not** count as copyright infringement. What, exactly, makes it onto that list varies from country to country, but common examples include:

- Research (like the Data Sitters Club)
- Private study
- Review or criticism
- Judicial proceedings or professional legal advice (lawyers look after their own!)
- Reporting the news
- Parody (and in some places, one or more of: caricature, pastiche, or satire) (lawyers like to laugh?)

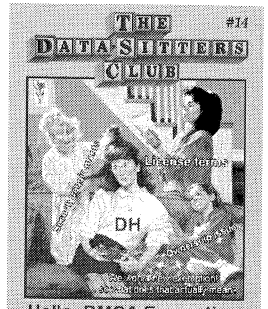
To make it as fair dealing, a use has to fit one of these purposes and be seen as "fair." So the list is a best-case scenario: it shows what might be allowed, if it meets a general test of fairness as the courts understand it.

https://datasittersclub.github.io/site/dsc7.html

Is my corpus limited to text? To print media? NO! But...

There is the Digital Millennium Copyright Act.

TDM = text data Mining



Hello, DMCA Exemption

Quinn Dombrowski, Rachael Samberg, Erik Stallman

Good news! We got a DMCA exemption granted that legitimizes breaking encryption so we can do TDM (which we know is a fair use). Bad news! Almost everywhere that sells the ebooks currently includes contractual language that prohibits us from making use of the DMCA exemption to break the DRM. Even worse news! Even if we did get a better contract, there's some super-difficult security requirements that I'm not even sure I could make happen at Stanford. Also, independent scholars and unaffiliated libraries and archives are shut out altogether. It feels like we've landed in the copyright law equivalent of middle school: in theory you have more freedom than before, but in practice, the experience of being there is pretty awkward and miserable (if you're not a fictional character in the Baby-Sitters Club series... and sometimes even if you are.)

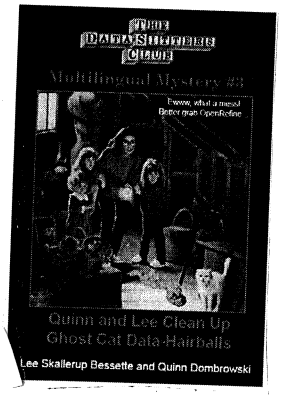
Make sure you are creating + using your corpus within the bounds of the law.

(Also, this is about the USA - Int'l laws vary).

https://datasittersclub.github.io/site/dsc14.html

Yup, ok, metadata is important... EXCEPT! What format is your metadata in?

This is where the inconsistencies arose, with each library offering its own idiosyncratic way of getting the metadata out of the library catalogue and into the hands of users. Quebec would only let me export a certain number of results from my search at a time as a csv file. I decided to limit my search by time period, as it was easiest to limit my searches in the interface that way. So, there are I had one CSV file for each year that the books were being published. France was the easiest to export, allowing me to download my entire search (which included the France and Belgium translations) as a giant CSV file. Spain let me export a page of results at a time as XML that displayed in my browser window, which I copied into a file. Shout-out to the Netherlands for having a functional English interface...and exporting the search as a simple text file. And then Italy...Italy made me cut and paste every single entry individually as XML as there was no batch export function.



(These Mystery covers don't greyscale well...)

Now, it's easy to go overboard with data cleaning. Sometimes my library degree gets the better of me, and I start organizing All The Things. (This in contrast to my usual inclination to just pile things in places and assume it'll all sort itself out when the time is right.) With DH in particular, there's a temptation to clean more than you need, and produce a beautiful data set that has all the information perfectly structured. It's usually driven by altruism towards your future self or other scholars. "Someday, *someone might want this.*" And if the data is easy and quick to clean, there's not much lost in that investment. But if it's a gnarly problem, and your future use case is hypothetical, maybe it's worth reconsidering whether it's the best use of your time *right now*. If someone wants a clean version of that data in the future, it's okay to let it be their problem. In the national library records that Lee found, there's a lot of data like that. Would it be interesting to see the publication cities for all the translations? Sure! Would it be fun to look at differences in subject metadata? Absolutely! (I'm really tempted by that one, truth be told.) But the data for both of those is kind of annoying to clean, and not actually what we need to answer the questions we're working on. Sometimes the hardest part of DH is setting aside all the things you *could do*, to finish the thing that you're *actually doing*.

https://datasittersclub.github.io/site/dscm3.html

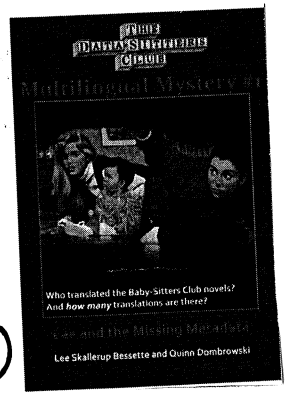
Talk to your metadata librarian and/or DH Specialist for proper guidance!!

BUT NOW...

You're going to tell me there's still MORE before we build this thing, aren't you? YES!!!

-METADATA-

(Trust me, you'll thank me later)



Metadata is relevant information about the individual pieces that will make up your corpus. Think: publication date + location, format, language, illustrator, etc... It could also be information relevant to your research question(s): Genre, main character, setting location, etc... Deciding what metadata you need is relevant is ESSENTIAL.

https://datasittersclub.github.io/site/dscm1.html

Have you carefully thought through what will go into your corpus?

Well, HAVE YOU???

https://datasittersclub.github.io/site/dsc19.html

"Where should people start with thinking about corpora, if they can't just get *everything*?" I asked.

Shelley thought for a moment. "The thing about corpora," she said, "is that there isn't a 'bad corpus' in the way that jokes can be bad. Usually, 'bad corpus' situations are ones where corpora are being used badly. Corpus construction is like making an argument[1], and the choices you make about how representative your texts are impact the kinds of conclusions you can make. That's true for all corpora. There are some good practices for how to get the best sample you can, and how to think about that in a principled way[2]. But first and foremost, you need to ask yourself, 'What are my goals with this project? What claims do I want to make?'"

"What if you're not really sure what exactly your research question is?" I asked. I'd felt this kind of chicken-and-egg unease before with YRDL – I wanted to do "something related to" this or that topic, but I didn't have a good enough sense of what even was there to know if any of my partially-baked ideas were going to go anywhere. I knew I'd have to actually put some thought into selecting a subset of YRDL than "everything that used X word at least Y times", but sometimes I'd start by just using AntConc or something to find all uses of the word in the entire collection of books, to see if it occurred frequently enough (and in a set of contexts that were at least vaguely interesting) to be worth pursuing.

"Honestly, a lot of it is more qualitative," said Shelley. "It really helps to *know* the corpus that you're working with. So if you're really at loose ends, go *read some things you're considering including*. The better you understand your materials, the easier it is to make well-informed choices about your research question, and what and how to select a good corpus."

"How do you know when you've got enough?" I asked. This was a corpus question that had plagued me before, like in DSC 11: Katia and the Sentiment Snobs when I was flailing around, grabbing dramatic middle-grade books with female protagonists to try to toss together a very ad-hoc and honestly pretty unthoughtful comparison corpus for the Baby-Sitters Club to try to figure out how emotive these kinds of texts tend to be.

"It depends on what kind of analysis you want to run," said Shelley. "It's easier for me because what I look at is more grammar than individual words – so there's always more occurrences of the things interesting to me in a text, than if you're looking for specific words or phrases. So I'd start by doing some poking around, using word searches or a concordance tool (like AntConc). How frequent is the word or phrase you're looking for? Corpus size does matter, but it matters more for some methods than others."

Do I have to start from scratch? From nothing?!?

No, but... Archives + datasets are PROBLEMATIC (like HathiTrust)

HathiTrust is massive: over 17.6 million volumes. Yeah, it's mostly English, but of the volumes with language metadata, English is just barely a majority at 51%. There's over half a million volumes in French, about a quarter-million each of Italian and Japanese, nearly 3,000 volumes of Macedonian, 17,000 volumes of Ukrainian, and almost 30,000 volumes in each of Thai and Turkish. When I think of HathiTrust, novels are what I imagine first, but they also have serials, journals, scholarly monographs, books in all subject areas, directories, legal documents – basically anything and everything you'd find in a large research library, because that's precisely where everything came from.

One of the nice things about the Data-Sitters Club is the way we've been able to mostly sidestep questions of corpus composition. What does our primary corpus have? All the *Baby-Sitters Club* books by Ann M. Martin. All of them. The Mysteries, the Super Specials, California Diaries, Friends Forever – if it has to do with Kristy Thomas and her friends, we have it. That's great for answering questions about that series (especially within the context of its own world-making), but for most computational text analysis and DH research, including any broader questions we want to ask to situate this series in different contexts, things aren't that simple. You need to put a lot of thought into how you construct your corpus, and it's very rare to be able to get all of **anything** the way we've gotten all the *Baby-Sitters Club* books. The points Roopsi makes in her book *New Digital Worlds* about the gaping holes, silences, and omissions in colonial archives shaped by the priorities and interests of empires also apply to HathiTrust. 17 million items feels so immense it's tempting to think it's got to be comprehensive. But you can't forget about how it came to be. HathiTrust is the recipient of data from the Google Books project, which scanned millions of books held by libraries at (mostly) US-based research institutions; from other mass-digitization projects like those of the Internet Archive (still active) and Microsoft (shuttered in 2008), as well as libraries' own local digitization workflows. The whole process was (and is) a bit of a hodge-podge, and there's tons of duplication . But before you start worrying about how Google decided to include this or that book, you shouldn't forget about another set of underlying selection factors that systematically disadvantage certain genres, such as romance, pulpy sci-fi, and children's and YA books: the books had to be held by a research university library. This has major implications, as we anticipated – and our Junior Officer Cadence quickly discovered.

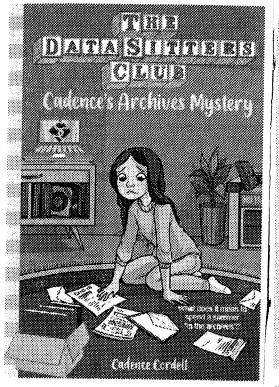
(Also, the interface isn't great so...)

Anyway, you have to think about what is and isn't in whatever corpus you use or create!

https://datasittersclub.github.io/site/dsc18.html

Can we start building the corpus Now? NOT YET!

Maybe you'll need to consult an archive (or 5) to build your corpus!



Here's the thing: depending on the size of the collection you're looking at and your familiarity with navigating finding aids and archival collections, just searching the finding aid for materials you might be interested in can be overwhelming. The Ann M. Martin Papers has 55 boxes in the collection — and while I was never going to look through all of them, the amount was enough to jolt my anxieties for a good couple of minutes before I managed to refocus and pick a few to start with. Depending on what your research is, the collection you're interested in might be relatively small or large. In both cases, the best way to prevent some early archival anxieties is to try and look at the finding aid and pick out materials to look at first before walking into the archives for the first time. Figure out what materials sound like they might apply to your research questions, list what boxes you are and aren't interested in, and get yourself organized. Reach out to the archivists on staff and talk to them about what you're researching and what specific collections you want to look at — they are the experts on the collections, and you shouldn't overlook them as resources. They certainly know what's in the archive better than you do. All of these methods should keep you from getting overwhelmed before you even set eyes on any archival sources (especially if your collection's finding aid isn't as detailed as that for the Ann M. Martin papers — but I'll come back to this a little later).

With all of these films and TV shows reflecting similar myths of archival research, you'd hope real life archival research is a similar experience. Unfortunately, the above scenarios are pretty unlikely. When it comes to archival research, you should expect more along the lines of these two scenarios:

a) there is not much in the archives in the first place; or

b) there's a TON of sources in the archives, and the vast majority of them are not related to your research question or interests.

More likely than not, there's not going to be something new and exciting you'll discover in the archives. And that's okay! Archival research doesn't have to reveal new and exciting information for your research to be valuable. But it also means that you shouldn't get too frustrated if you can't find what you were hoping for. Archives, in general, are a mixed bag — sometimes they have exactly the sort of materials you were hoping for, and sometimes they don't hold much at all. What's important is to manage your expectations, be flexible, and (if your collection is on the larger side)...

Roopsi has... issues w/ this :)

Get ready to do a LOT of skimming.

Skimming is probably the most important skill you'll need to develop while researching archival collections. No joking. And while this is partly due to not everything in the collection applying to your research questions, it's also due to time constraints on your research. I'm lucky enough that my internship is about two months long — plenty of time for me to work through the Ann M. Martin papers without stressing about getting to everything the DSC is interested in. Most researchers aren't that lucky. Depending on the funding involved and what the academic calendar for your institution looks like, it's more likely that your research period will be between a week or two to a month. Skimming will be a necessity, if only so you can sort through all the materials and have ample time to focus on the sources that do matter for your research. And, secondly, well, you remember that assumption we made earlier that the finding aid is 1) accurate and 2) detailed, and 3) generally is the best way to give you a sense of what "might" be in the archives?

https://datasittersclub.github.io/site/dsc17.html