How many OER are there?

Jonathan A. Poritz

jonathan@poritz.net
poritz.net/jonathan

20 October 2022
Open Education Conference 2022

Intro: Land acknowledgement

Before we begin, I must acknowledge that I began this work while I was living in the unceded territory of the Ute Peoples. The earliest documented people in the area also include the Apache, Arapaho, Comanche, and Cheyenne. An extended list of tribes with a legacy of occupation there can be found on the Colorado Tribal Acknowledgement List.

I am grateful for the chance to have lived and worked in that beautiful place and will always cherish that memory, even though I am no longer a resident there.[1]

[1] Where I live now there is no tradition of land acknowledgements of which I am aware.

The question, and what to do about it

A while ago, I wondered

How many OER are there?

[Hence the title of this talk.]

In this presentation, I will tell you about how I tried

- to decide what kind of answer I would be happy with,
- to make the question a bit more precise,
- to go about getting that answer, and
- to understand what answer I was actually able to get.

What kind of answer do I want?

Whenever I see a single statistic, I feel like it is begging for some *context*.

I also did not specify, when asking the question, a particular date and time. Since the number of OER is probably constantly changing – *growing*, one imagines – the best thing might be to give an answer for all possible times one might specify.

In fact, let's put all these numbers together in a *graph* and say that

I want a graph of how many OER there have been, over time.

In fact, this is closer to what I was originally wondering: I wanted to write a sentence about the well-known (presumably) trend of growth – exponential, maybe? something like that – of the body of existing OER. But I couldn't find any prior results on the topic![2] Hence this project....

[2] I did ask smart people! *E.g.,* Nicole Allen responded, and she suggested that this wasn't actually the right question, that a better question would be about *engagement* with OER. I absolutely agree! But my less important question is still of some interest to a data geek like me.

Making the question precise: What are those "OER?"

What are the things I want to count (repeatedly, at different times, to make a graph)?

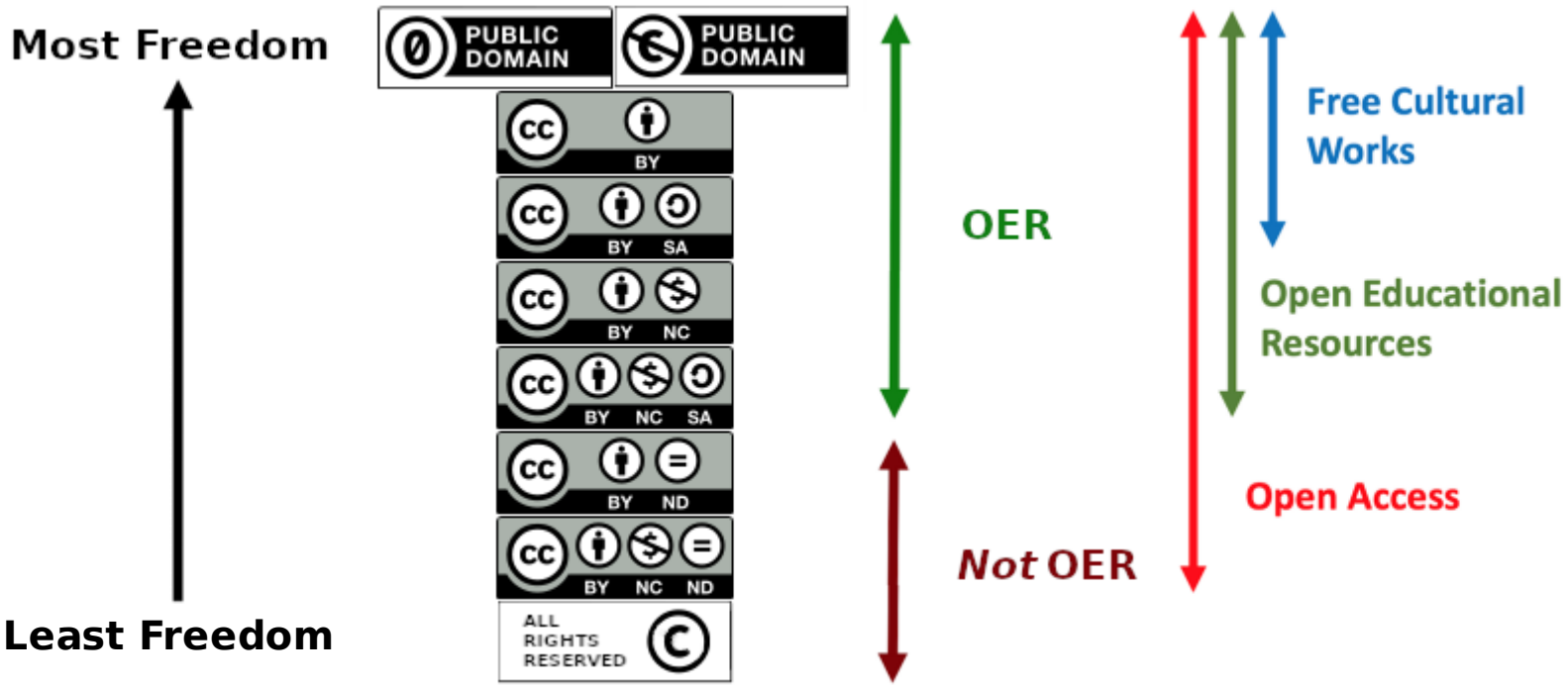
They are "Open Educational Resources [OER]."

Fortunately, the UNESCO OER Recommendation can be taken as canonical:

> "1. *Open Educational Resources (OER) are learning, teaching and research materials in any format and medium that reside in the public domain or are under copyright that have been released under an open license, that permit no-cost access, re-use, re-purpose, adaptation and redistribution by others.*
>
> 2. *Open license refers to a license that respects the intellectual property rights of the copyright owner and provides permissions granting the public the rights to access, re-use, repurpose, adapt and redistribute educational materials."*

Making the question precise: UNESCO and *licenses*

Mapping the UNESCO definition to the Creative Commons [CC] licenses and public domain tools (the most common approaches for OER), we get:



In particular, then,

The OER I want to count must all bear a CC PDM or CC0 tool or a CC BY, BY-SA, BY-NC, or BY-NC-SA license.

Caveat: There could be other licenses or copyright statuses.

There are other licenses which might meet the criteria expressed in the UNESCO definition of OER! The Open Textbook Library [OTL] from the Open Education Network [OEN] has seventeen works which bear the GNU Free Documentation License, which seems to meet the UNESCO OER definition.

Other works might have fallen into the global public domain but not bear the CC PDM, simply because no competent authority had bothered to put one on a commonly accessed version of the work. These should nevertheless counted amount OER.

Finally, the Creative Commons does not recommend that its licenses be used for *software*, saying there are many others which are better adapted to the particular needs of code: see, *e.g.*, a list of approved open-source licenses from the Open Source Initiative. Since more and more OER – even ones which are basically "textbooks" – may incorporate (as interactive elements) or be nearly entirely (as in Jupyter Notebooks or similar) *code*, the open education community probably needs to stop using slides like the one on the last slide[3] which portray OER as ***necessarily*** carrying one of those CC licenses/statuses.

[3] Yes, I am criticizing myself!

Making the question precise: UNESCO and *materials*

UNESCO says that OER are are *"...learning, teaching and research* ***materials****"* [emphasis added].

These could be classroom handouts, test banks, individual diagrams, videos, pieces of software, *etc.*

Amorphous materials like that are hard to count, except perhaps as pages or megabytes, which I will not do.

One can, presumably, count *books*, though. So

The OER I want to count should all be "textbooks."

There is certainly a tradition of doing this in the open education space. *E.g.*, before the OEN broadened its scope to all of Open Education, it started out as the **Open Textbook Network [OTN]**! We may eventually count more significant things like engagement, but shouldn't we at least start by counting textbooks?

Caveat: What exactly is a "textbook"?

There has been some interesting discussion in recent years about "what is a 'book' in the age of the web?," what will be "the textbook of the future"[4], and even if textbooks are the best tools for learning, even for courses that have traditionally used them[5].

But surely at least "traditional textbooks" have a clear definition?

A number of organizations who run open educational professional development programs or who otherwise support the movement, including the Rebus Community and the OEN, often speak about textbooks as being in some tension with books that might be called "academic monographs": textbooks have to have some fairly fixed structure and common features like chapter openers and closers, pedagogical elements, exercises or discussion prompts, learning outcomes, *etc.*

In my own education, I took great courses which used academic monographs (or novels or other forms of books). So I will make a very informal definition of "textbook" as anything which its author or some collector or cataloger has called a textbook. Many "monographs" that might be used for education will therefore be accepted in my OER counting project, for example.

[4] Just use that phrase with your favorite search engine!

[5] See, *e.g.*, my talk at the last in-person Open Ed conference, on "Just-In-Time Educational Resources".

Making the question precise: What will time be in the graph?

If I am going to count textbooks with certain legal statuses over time to make a graph, I need to know what is meant by "time" in this project.

My original motivation was to make that time plot of the number of OER, with an idea to showing how many OER "have been available to the community" over time. So I really want to know when these textbooks have been made public.

Often this *publication date* will be the same thing as the work's *copyright year* – copyright springs into existence in the US when a work is "fixed in a tangible medium of expression"[6] – and I suspect that most folks who go to the trouble of creating or adapting a OER do so with the intention of using it, so they make it public just about as soon as it is created [fixed, in the US].

Therefore,

"Time" in my graph of the number of OER will be the publication date or copyright year, whichever is possible to determine for each particular OER.

[6] and in most other countries when the work merely created, even without fixation.

Making the question precise: When should I count OER as different?

To count things, you must know when two to be considered as the same or different, in order to know when to increase the count.

Since a foundational value, and commonly followed practice, in open education is to *adapt* existing works, similar OER abound!

Fortunately, this is a problem which has already been solved in copyright law, where often one must ask if two works are enough "the same" for one to count as a copy of the other.

That means

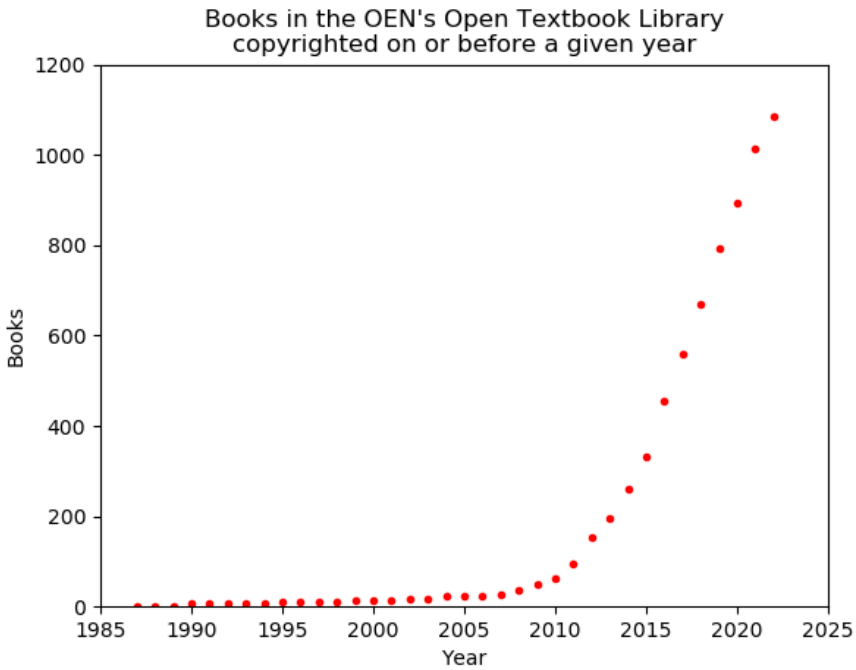
I will count books which copyright law would consider all different.

In particular, this means that a printout of an ebook is not a new book, nor is an electronic version which is in a different file format from the original, nor is a version which fixes a few typos or changes a font.

A translation of a book, however, will almost always be considered a new work, as will essentially any new infusion of original authorship.

A test case: the OEN's OTL

How about using a limited but high quality dataset to see if the approach described above makes sense.
In particular, the OEN's OTL definitely consists of textbooks, and the OEN shares a catalog CSV which tells the works' licenses and copyright years. Removing entries for books which do not have the correct license or copyright status, extracting the copyright years, and making the resulting graph, gives this:



for this:

The OTL's catalog CSV was downloaded to a local copy OTL.csv.

Rows corresponding to items with incorrect licenses were removed, the column of copyright years was extracted and sorted into a file otl_copyright_years.

The graph was produced using the command

```
./cyears_graph.py -d otl_copyright_years -r
  "the OEN's Open Textbook Library" -s 1985
  -t 1200 --endyear 2025
```

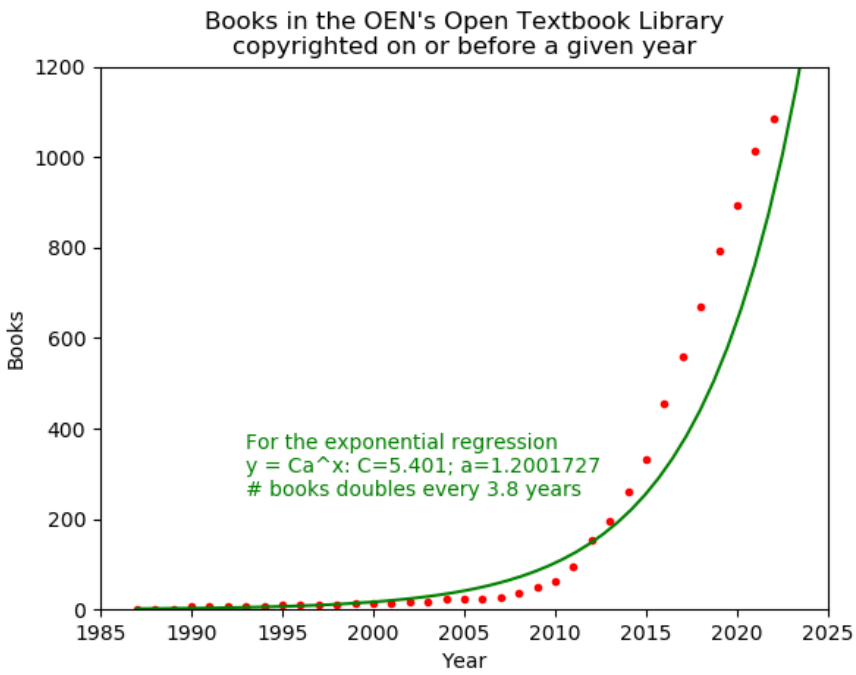
(all on one line)

This used a(n openly licensed (of course)) Python script cyears_graph.py

Exponential fitting to the OTL graph

I was hoping, you may recall, that the graph of how many OER there are would show something like an exponential growth, and that scatterplot does really seem to take off.

Unfortunately, the best exponential fit is not very good:



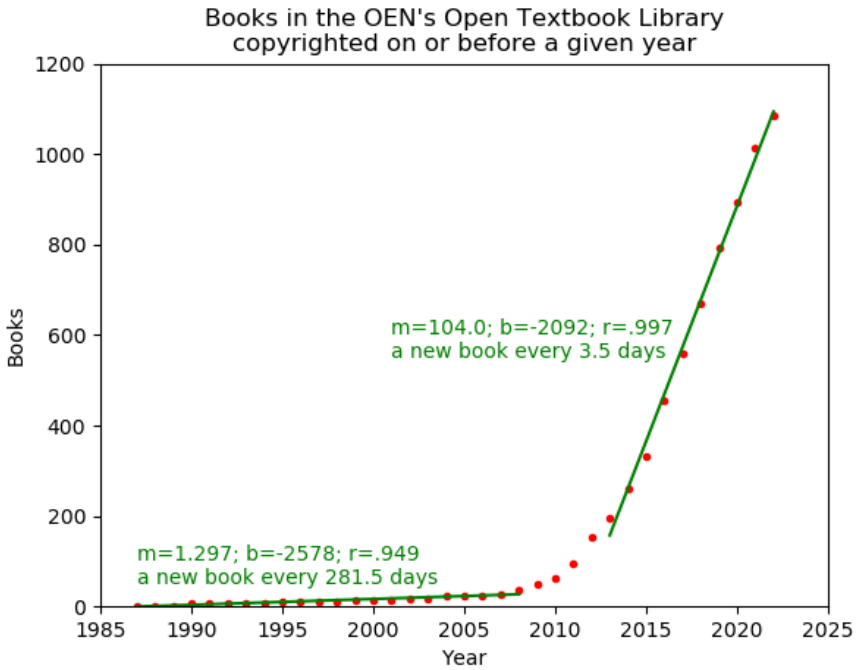
command used here:

```
./cyears_graph.py -d otl_copyright_years -r
   "the OEN's Open Textbook Library" -s 1985
   -t 1200 -e '(1987,2024):(1993,250)'
   --endyear 2025
```

(all on one line)

Piecewise linear fitting to the OTL graph

Looking at the original scatterplot, it in fact seems as if there are two quite linear regimes:



command used here:

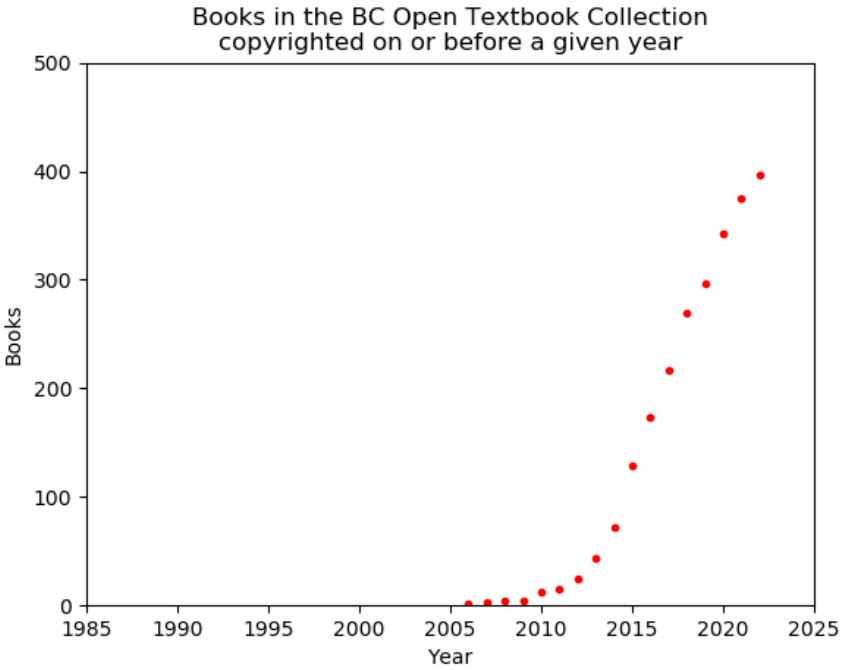
```
./cyears_graph.py -d otl_copyright_years -r
   "the OEN's Open Textbook Library" -s 1985
   -t 1200 -l '(1987,2008):(1987,50)'
   -l '(2013,2022):(2001,550)'
   --endyear 2025
```

(all on one line)

Trust a data analyst: linearity is very rare in nature. I would guess that during the whole life of the OTL, there have been too many new OER to be processed. Instead, in each of the two different linear regimes in this graph, there were two different systems or groups of personnel who had a fixed rate (different between the two regimes) of ingesting a certain number of new OER per day, and they just always operated at capacity.

Another test case: the B.C. Open Textbook Collection

Another limited but high quality dataset is provided by the B.C. Open Textbook Collection from BCcampus, consisting of resources which all have good OER licenses and which again are all clearly textbooks.



for this:

The "Full Set (.mrc)" of MARC records was downloaded from the page BC Open Textbook MARC Records to a local copy BCopentextbooks_RDA_fullset_Q2_June30_2022.mrc.

Using a tool marc2excel, this was converted to the file BCopentextbooks_RDA_fullset_Q2_June30_2022.xlsx whose column of copyright dates was extracted and sorted into a file BCcampus_copyright_years.

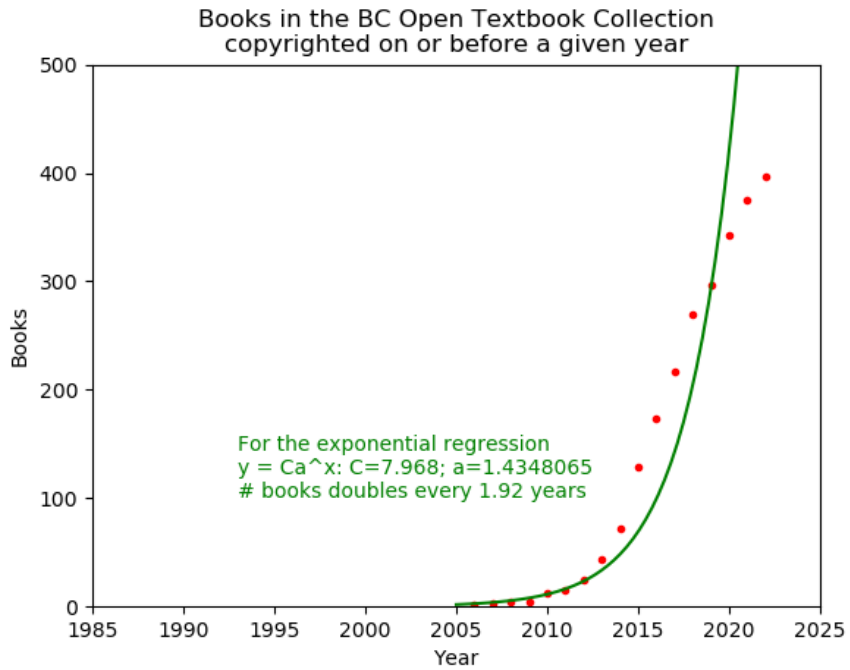
The graph was produced using the command

```
./cyears_graph.py -d BCcampus_copyright_years -r
  "the BC Open Textbook Collection" -s 1985
  -t 500 --endyear 2025
```

(all on one line)

Exponential fitting to the BCcampus graph

Here again, the best exponential fit is not very good:



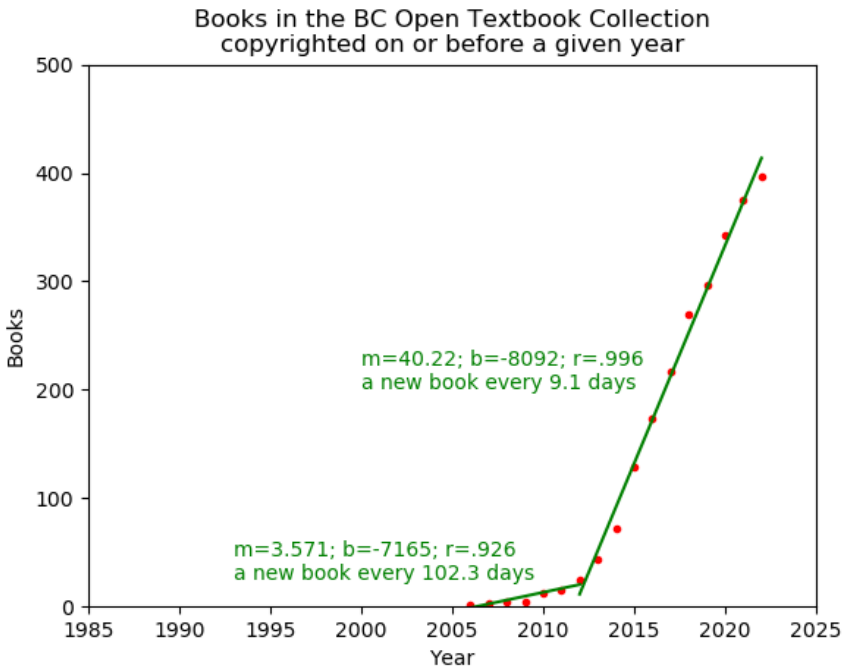
command used here:

```
./cyears_graph.py -d BCcampus_copyright_years -r
   "the BC Open Textbook Collection" -s 1985
   -t 500 -e '(2005,2022):(1993,100)'
   --endyear 2025
```

(all on one line)

Piecewise linear fitting to the BCcampus graph

Once again, looking at the original BCcampus scatterplot, it in fact seems as if there are two quite linear regimes:



command used here:

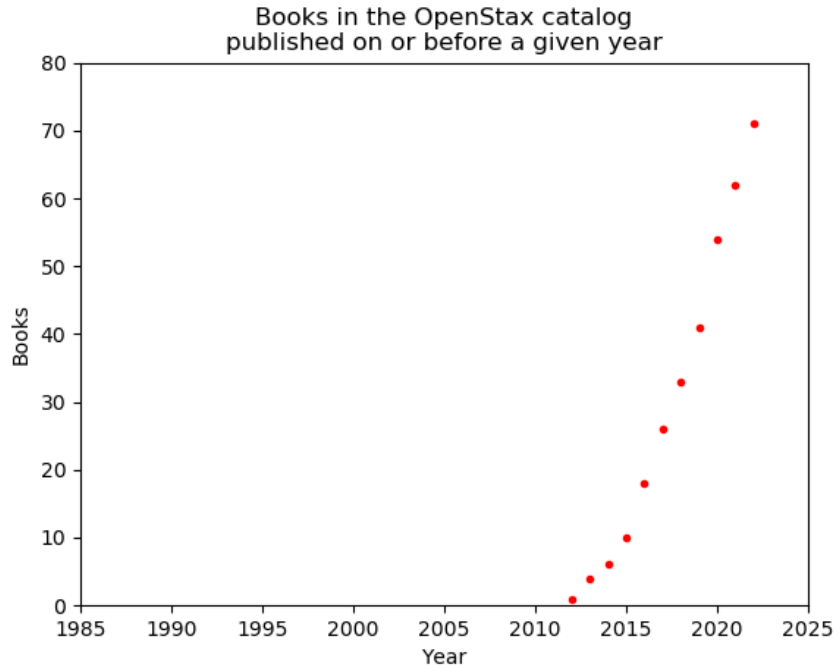
```
./cyears_graph.py -d BCcampus_copyright_years -r
   "the BC Open Textbook Collection" -s 1985
   -t 500 -l '(2012,2022):(2000,200)'
   -l '(2006,2012):(1993,25)'
   --endyear 2025
```

(all on one line)

The same data analyst's hypothesis about what is causing the "hockey-stick" shape of this graph apply here as in the OTL case.

Another test case: OpenStax

Another limited but high quality dataset is provided by the textbooks from OpenStax. While these are all clearly textbooks, and there are few enough that it is easy to look at their descriptions one by one to find the relevant information, OpenStax reporting of dates is a bit odd. For each of their books, they give both a "publication date," often a few years ago, and a "copyright date," often in the last year or so. Why these are different is unclear to me. For the reasons described above, I used the publication date,



publication years extracted from each book's description on OpenStax's website and sorted into a file OpenStax_pyears.

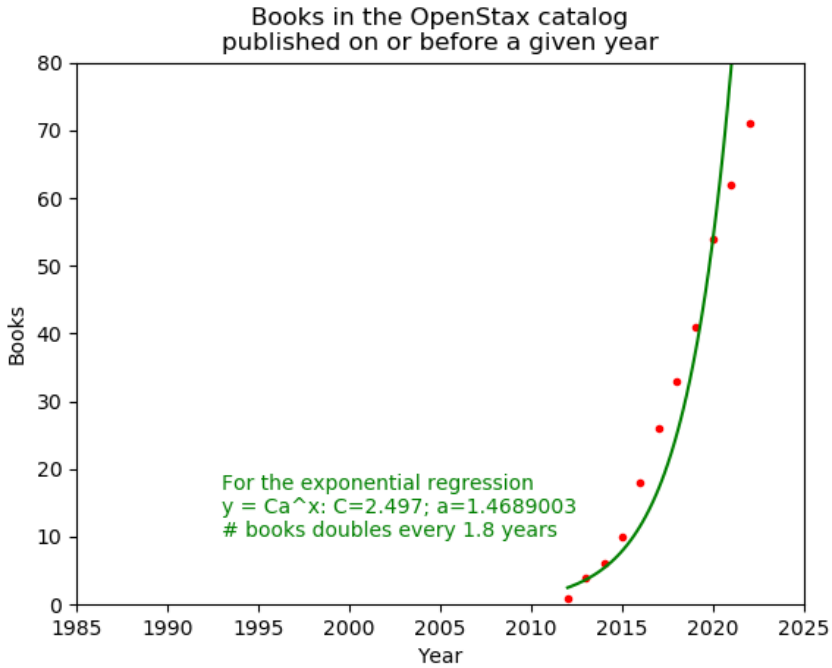
then command used was:

```
./cyears_graph.py -d OpenStax_pyears -r
   "the OpenStax catalog" -s 1985 -t 80
   -v "published" --endyear 2025
```

(all on one line)

Exponential fitting to the OpenStax graph

Here again, the best exponential fit is not very good:



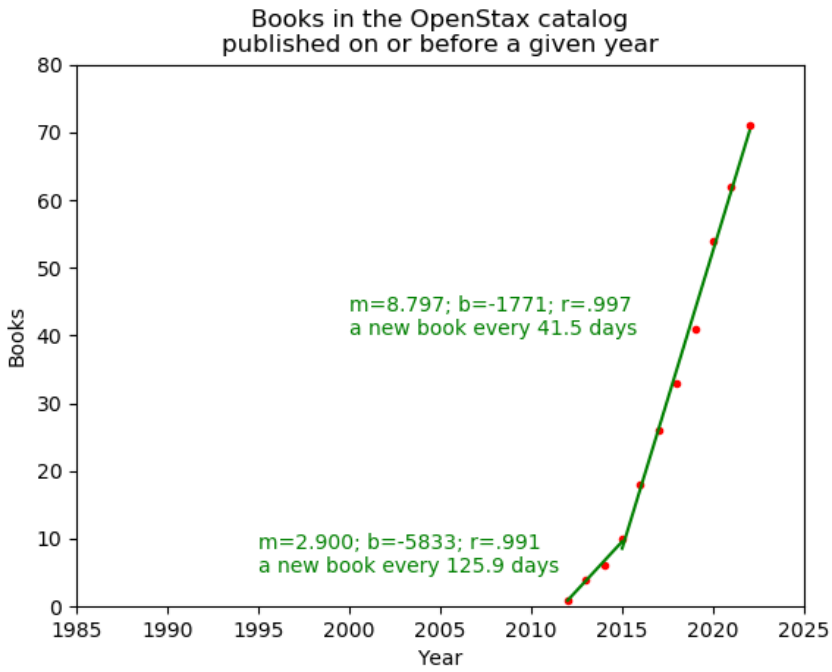
command used here:

```
./cyears_graph.py -d OpenStax_pyears -r
   "the BC Open Textbook Collection" -s 1985
   -t 80 -e '(2012,2022):(1993,10)'
   --endyear 2025 -t 80 -v "published"
```

(all on one line)

Piecewise linear fitting to the OpenStax graph

Once again, looking at the original OpenStax scatterplot, it in fact seems as if there are two quite linear regimes:



command used here:

```
./cyears_graph.py -d OpenStax_pyears -r
   "the OpenStax catalog" -s 1985
   -t 80 -l '(2012,2015):(1995,5)'
   -l '(2015,2022):(2000,40)'
   -v "published" --endyear 2025
```

(all on one line)

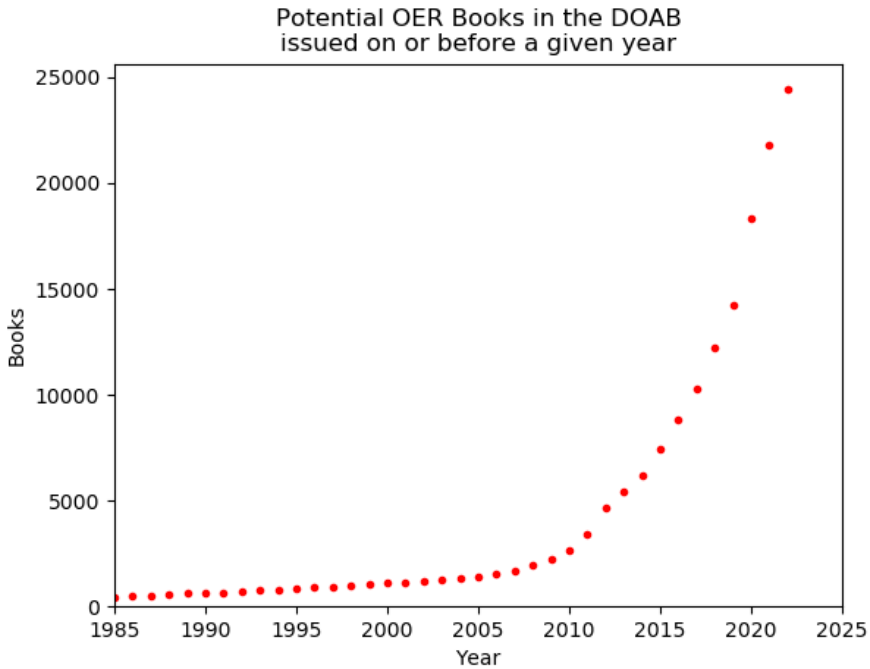
The same data analyst's hypothesis about what is causing the "hockey-stick" shape of this graph apply here as in the OTL case.

Another test case: The Directory of Open Access Books

The Directory of Open Access Books [DOAB] is a large and wonderful site containing OA books (and some other resources, which I will filter out) from a number of sources.

For reasons given above, I will count DOAB books, when licensed correctly, as OER, even though many look like academic monographs rather than textbooks.

The DOAB catalog, available from their page Metadata for Libraries and Aggregators as the file repository-export.csv gives good information on the date the works were "issued" – which I will take as the publication date – and the licenses, from which we can filter for the UNESCO OER definition-compatible ones. I did this in a Python script doab.py.



output of doab.py script put in file DOAB_iyears

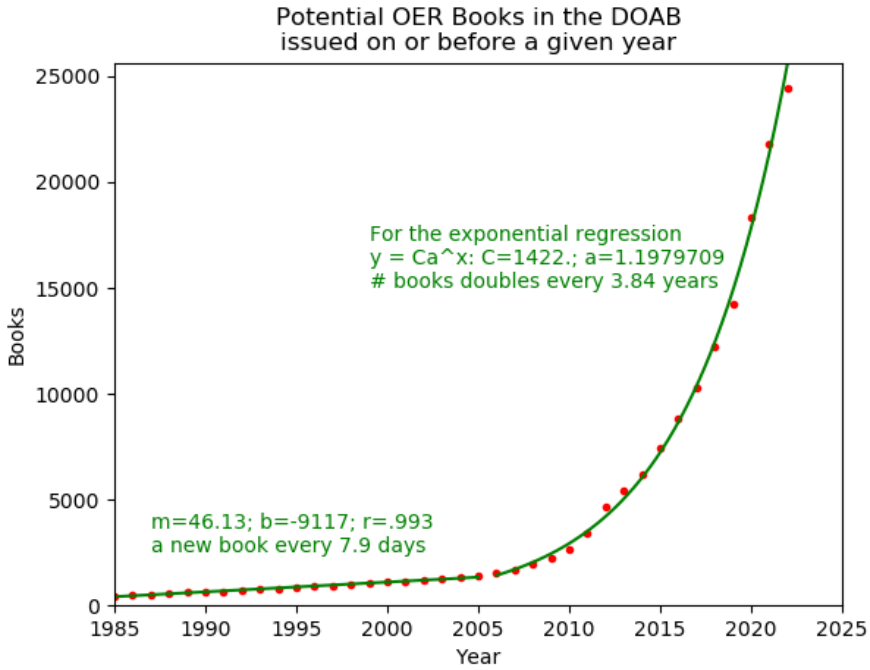
then command used was:

```
./cyears_graph.py -d DOAB_iyears -r
  "the DOAB" -s 1985 -v "issued"
  --endyear 2025 -p "Potential OER "
```

(all on one line)

Combined linear and Exponential fitting to the DOAB graph

The DOAB seems to have a regime with good linear fit and then one with good exponential fit, that's fun!



command used here:

```
./cyears_graph.py -d DOAB_iyears -r
   "the DOAB" -s 1985 -v "issued"
   -l '(1985,2005):(1987,2500)'
   -e '(2006,2022):(1999,15000)'
   --endyear 2025 -p "Potential OER "
```

(all on one line)

Let's go!

OK, it's pretty obvious how to wrap up the whole thing:

1. Put together a list of all existing OER.
2. Remove from the list all items that are not "textbooks."
3. Remove from the list all items that do not have the Creative Commons licenses or copyright statuses we are permitting.
4. Make another list, consisting of all of the publication or copyright years of each of the items remaining on the first list.
5. Sort the list of dates, count how many dates are in each past year, and make the corresponding graph.

Oh, snap.

I think you've seen the flaws in my approach.

How are we going to get "a list of all existing OER?"

Any approach which simply crawls aggregates many known OER repositories will both miss enormous numbers of useful OER and also catch some OER multiple times.

(In principle removing duplicates can be done by hand or with code, but it will be hard to know that "Calculus, Second Edition" and "Calculus 2e" are the same thing, and checking whether enough new creativity has been added to make a different version of a preexisting book an adaptation and not merely a copy will be **very hard**.)

I will try to continue aggregating repository catalogs and doing my best to make an exhaustive list without duplicates, but this is an huge task that will not have good results particularly soon ... and will probably never be completely finished. Of course, any partial answer will still give a lower bound on the number of OER that exist, so that is some good information.

What **have** we learned? (specifically)

I think that the data analyst's perspectives on the graphs above lead us to conclude that, essentially, the body of available OER is growing exponentially with a doubling time of about 3.8 years.

Or, at least, it wants to grow exponentially, at the moment[7]. It seems that in many particular organizations, though, the available capacity is limiting the growth to be linear, with growth often on the order of one new OER shared every few days.

This suggests that, in the short term, adding capacity to support groups polishing and sharing OER will result in greater output and greater numbers of OER available to the community, almost without bound, at present. At some point we will will hit the end of that exponential growth, but that will likely be in the saturated environment where just about everything is already and always OER – that's a world I'd be happy to live in, even if the growth curve then levels off to something linear!

[7] Data analysts also will say that exponential growth is common in nature ... but only for short periods of time, before the environment gets saturated.

What **have** we learned? (more generally)

OER are spread out all over the Internet, so it is *very hard* to do research on them. Of course, we already knew this. It's also not necessarily a bad thing that they are so spread out – a prominent figure in the OER world said a few years ago that they wanted their site to be the "Facebook of OER." (This was before Facebook's recent loses of market share.) I'm not really happy with that vision, TBH, even though it would make this current research project much easier.

Many OER folks aren't very careful to publish clear metadata, with licenses and copyright/publication dates clearly shown. This has the potential to be a big problem in the future, and certainly makes the 5Rs more difficult in an entirely unnecessary way! So: Please mind your metadata, OER folks!

Clear metadata that is easily findable (in standard places and in standard formats) will enable research like this project to work by simply crawling the web and harvesting this metadata. Since we know from the whole history of the Internet that crawling the web and looking for the things we want works much better than trying to have curated lists of "good stuff" on the 'net, probably that approach will work better also in the OER space, if only there is good metadata. So I encourage my wonderful librarian colleagues in open education not to try to make catalogs and all-encompassing repositories, but rather to concentrate on helping the community make good metadata easy and the norm.

Thanks

I'd like to thank the following folks for sharing with me their data and, more importantly, their insights:

- Lauri Aesoph
- Nicole Allen
- Amanda Coolidge
- David Ernst
- Josie Gray
- Delmar Larsen
- Karen Lauritsen
- Ethan Turner
- Steel Wagstaff

[In no order other than the arbitrary one determined by the alphabet and their last name[8].]

Thank you all so much!

[8] Sorry, Steel.

Discussion and contact info

Discussion!!

Contact info:

Email: jonathan@poritz.net ; Tweety-bird: @poritzj .

Get these slides at poritz.net/j/share/howmanyOER.pdf and all files for remixing[9] at poritz.net/j/share/howmanyOER/ .

If you don't want to write down that full URL, just remember
- poritz.net/jonathan/share
- or poritz.net/j/share
- or poritz.net/jonathan [then click **Always SHARE**]
- or poritz.net/j [then click **Always SHARE**]
- or scan → [then click **Always SHARE**]

[9] subject to CC-BY-SA 4.0