

1	Introduction
1.1	Economic Pressures on Monograph Publishing
1.2	Reviewing the Evidence
2	Accessible Data on Open Monographs
3	Explorations
4	Future Puzzles ...
5	About this Document
	References

Exploring the Public Evidence on Open Access Monographs

Code ▾

Micah Altman

MIT Libraries – Center for Research on Equitable and Open Scholarship (<https://libraries.mit.edu/creos/>)

2021-01-14

1 Introduction

There is ongoing tension between the desire of scholars to share their work widely and openly, and the need to fund the infrastructure and labor of publishing. One place in which this tension is most evident is in the sale of scholarly monographs. While they are only a small fraction of scholarly communications volume, market, and readership – academic monographs continue to play an important role in the humanities and social sciences. They represent an important form of long-form scholarship – not readily expressible through journal-length publications. And, as such, monograph publication through a university press forms a critical component of tenure evaluation – sometimes independent of the extent to which the monograph itself is purchased, read, or cited. (Eve 2014; Crossick 2016)

Die
M i t s c h u l d i g e n
Ein Lustspiel.

von

G o e t h e.

A d t e A u s g a b e.

Leipzig,
bey Georg Joachim Göschen,
1787.

First Page from the Oldest Open Monograph

1.1 Economic Pressures on Monograph Publishing

Monograph publication has been in crisis for approximately two decades. Changes in academic library collection policies — driven, in part, by the serials crisis and the greater integration of purchase-on-demand workflows — have led to traditional monograph publishing becoming generally unprofitable. (Crow, n.d.; Spence 2018) At the same time, there is an increasing

demand among scholars, research funders, and the public that the outputs of scholarship be made open access. (Guédon 2019; Science Europe, n.d.)

There are many potential funding models for open monographs (Penier, Izabella, Eve, Martin Paul, and Grady, Tom 2020; Adema, Stone, and Keene, n.d.). Currently, a number of initiatives seek to promote consortial models involving both publishers and groups of libraries. These consortial models include library crowdfunding, membership fees, subscribe-to-open transition, and the direct funding of shared infrastructure. These models act to coordinate disciplinary communities (usually through libraries as representatives); enable publishers to streamline workflows for open digital publication; and reduce potential cost-risk (to publisher and reader).

These initiatives notwithstanding, open access monographs constitute a small fraction of the total monograph titles now and in the near future, and will likely make up a few percent of monograph titles published annually. (Grimme et al. 2019)

1.2 Reviewing the Evidence

Open monograph publishing remains in its early stages. The CREOS (<https://libraries.mit.edu/creos/>) “The Economics of Scholarly Monographs” project is an examination of this area. This fall, as an initial step, we published an annotated bibliography (<https://libraries.mit.edu/creos/research/economics-of-scholarly-monographs/>) that serves as a jumping off point for scholars to explore the effects of open availability on monograph revenues.

In this blog post we look at the open data available on monograph publication, and use it to explore patterns and trends in open monograph publishing. This blog post takes the form of a guided, interactive, reproducible data analysis based on currently available public data.¹ We aim for this exploration to inform libraries, publishers, and authors about the landscape, and prepare for future transitions to open access.

2 Accessible Data on Open Monographs

The most complete index of open access monographs is the Directory of Open Access Books (<https://www.doabooks.org/>) (DOAB), which lists tens of thousands of individual monographs (also known as ‘titles’). DOAB makes its metadata index available as open data.

[Hide](#)

```
# core libraries for tidy data science in R
library(tidyverse)
library(magrittr)
if (doc_debug) {
  require(tidylog)
}

## the details of data retrieval in a separate module, included in our repository
## mono_load_* loads the named data as a R data frame from cache in github
## mono_fetch_* routines are used to retrieve a new version of data from canonical source

source("fetch_data.R")

## ISBN normalization and retrieval of open descriptive metadata based on
## these are implemented through the isbnTools python module
## we install these and provide a simple R wrapper (based on reticulate)
source("isbnTools.R")

## Helper functions for data visualization
source("plotly_helper.R")

if (doc_refresh_data) {
  isbn_tools_init()
  mono_fetch_doab()
  mono_fetch_oapc()
}
doab_df <- mono_load_doab()
oapc_df <- mono_load_oapc()
```

The unique identifiers in the DOAB can be used to link it with other data sources. As an example, we can use the ISBN as a key to retrieve information from [Google Books](#). For example, we can retrieve and display the cover of the most recently added title:

[Hide](#)

```
latest_book_isbns <- doab_df %>%
  arrange(`Added on date`) %>%
  ungroup %>% slice_tail() %>%
  select(`ISBN`) %>%
  str_split(" ") %>% unlist() %>% as.character()

if (doc_refresh_data) {
  cover_uri <- isbnTools("cover", latest_book_isbns[1])[1,]
  # increase zoom level
  cover_uri %>% str_replace("zoom=5", "zoom=10")
  download.file(cover_uri, doc_sample_thumbnail_path)
}
```



Karrierewege zur Professur an einer Fachhochschule

**Ursula Diallo-Ruschhaupt
Susanne Plaumann
Eva-Maria Dombrowski**

**Schriftenreihe „Gender-Diskurs“
des Gender- und Technik-Zentrum (GuTZ)
der Beuth Hochschule für Technik Berlin
Band 09
Herausgeberinnen
Eva-Maria Dombrowski, Antje Ducki**



Cover of Latest Monograph, Retrieved from Google Books



The DOAB data also provides links to the text of the open monograph itself. The monograph content is thus potentially available for harvesting, analysis, and integration with other sources. In practice, however, retrieving the content through DOAB may require some additional web scraping, as demonstrated below. For books also available in [Hathitrust](#) obtaining the content through their APIs is more reliable and straightforward.

[Hide](#)

```
### Capture image of first page of oldest open monograph
library(rvest)

## find the oldest book in DOAB and extract its URL
oldbook_url <- doab_df %>%
  arrange(`Year of publication`) %>%
  head(n = 1L) %>%
  select(`Full text`) %>%
  as.character()

if (doc_refresh_data) {
  ## retrieve book page follow metadata embedded in webpage
  require(rvest)
  oldbook_pg <- read_html(oldbook_url)
  pdf_url <- oldbook_pg %>%
    html_nodes(xpath = '//meta[@name="citation_pdf_url"]') %>%
    html_attr("content")

  ## retrieve book and extract first page as image
  require(pdftools)
  pdf_tmpfile <- tempfile(fileext=".pdf")

  download.file(pdf_url, pdf_tmpfile)
  pdf_convert(pdf_tmpfile, page = 1, dpi = 300, file = doc_sample_image_path)
}
```

Two other data sources are designed to provide additional information specifically about open access monograph titles:

- The OpenAPC (<https://www.intact-project.org/openapc/>) project provides title-level data on processing charges, supplied by a number of consortial initiatives.
- OpenBookPublishers (<https://www.openbookpublishers.com/section/92/1>) provides title-level usage data on the titles it publishes.

In addition there are a number of publicly accessible (not necessarily open) sources of metadata about large collections of books generally. The most notable comprise:

- *Descriptive Metadata*: ISBN registries including the service provided by OpenLibrary (<https://openlibrary.org/>) can be used to obtain additional descriptive metadata for titles, including subject headings. The open ISBNtools (<https://pypi.org/project/isbntools/>) package provides a standardized way of retrieving this data from a range of registries.
- *Citations*: A limited number of monographs are assigned DOI's indexed in [CrossRef](#), open citation data is available through the I4OC initiative (<https://i4oc.org/#faqs>). Commercial services such as [Google Scholar](#), [Dimensions](#), and [Scopus](#), also include some citation information for selected books. This information is challenging to access systematically, but small collections can be extracted using Harzing's Publish or Perish (<https://harzing.com/resources/publish-or-perish>) tool.
- *Public domain works*. A range of books, including some monographs, are now open by virtue of coming out-of-copyright and into the open domain. These are not listed in DOAB – however API's for HathiTrust (<https://www.hathitrust.org/data>) and JSTOR (<https://www.jstor.org/dfr>) provide descriptive metadata, rights metadata, and text-analytic metdata (e.g. ngrams) for the (open) books in their collection.
- *Prices*: Amazon provides pricing API's (<https://webservices.amazon.com/paapi5/documentation/use-cases/buying-price.html>) that can be applied to monograph titles, and a number of third parties track Amazon price histories. This data is available under restrictive terms, and in small quantities.

3 Explorations

In the table below you can browse a sample of titles:

[Hide](#)

```

library(DT)
## interactive sample data table
doab_df %>%
  ungroup() %>% slice_head(n = 1000) %>%
  datatable(class = "cell-border stripe", caption = "Sample of DOAB Catalog",
            options = list(pageLength = 5), extensions = "Responsive")

```

Show **5** entries

Search:

Sample of DOAB Catalog

	Title	ISBN	Volume	Authors	Pages	ISSN	Series title
+	1 Die	9783839445655		Mohan, Robin	344		Sozialtheorie
+	2 Ökologische und ökonomische Bewertung von Co-Vergärungsanlagen und deren Standortwahl	9783866443556		Koch, Matthias	244 p.		
+	3 Die	9783839407530		Hofer, Stefan	322		Lettre
+	4 Estudios de intertextualidad semítica noroccidental. hebreo y ugarítico	9788491682431		Gregorio del Olmo Lete	526		Collecció Barcino Monographica Orientalia
+	5 La Adopción y el acogimiento: presente y perspectivas	9788491682066		Diana Marre ; Joan Bestard	342		Col·lecció Estudis d'Antropologia Social i Cultural

Showing 1 to 5 of 1,000 entries

Previous 1 2 3 4 5 ... 200 Next

3.1 Dirty Data: Lather, Rinse, Repeat.

After browsing the DOAB sample for a short while, you will likely notice glitches. There are many, including missing fields, typos; undocumented and inconsistent formats for names, dates, and identifiers; and multiple values packed into a single field in undocumented and inconsistent ways. These ‘dirty data’ issues are not unique to DOAB, and are in fact, ubiquitous across the data sources we examined. For further data integration, at minimum, standardization of date and ISBN fields is required, as illustrated in the code below.²

Hide

```

library(lubridate)
### Data Cleaning
## address basic issues with:
## - date standardization
## - ISBN list packing
## - ISBN format standardization
## - non-monograph entries

## DOAB basic data cleaning
doab_df %>%
  filter(`Type` == "book") %>%
  mutate(
    DT_PUBLISHED_YR = year(parse_date_time(`Year of publication`, "y")),
    DT_ADDED_YR = year(parse_date_time(`Added on date`, "ymd HMS")),
    LS_ID_ISBNS = str_split(
      str_replace_all(ISBN, "[^0-9\\s]X*", ""), "\\s+"
    ) %>%
    mutate(LS_ID_ISBNS = lapply(LS_ID_ISBNS, isbntools, meth="ean13"))
  )

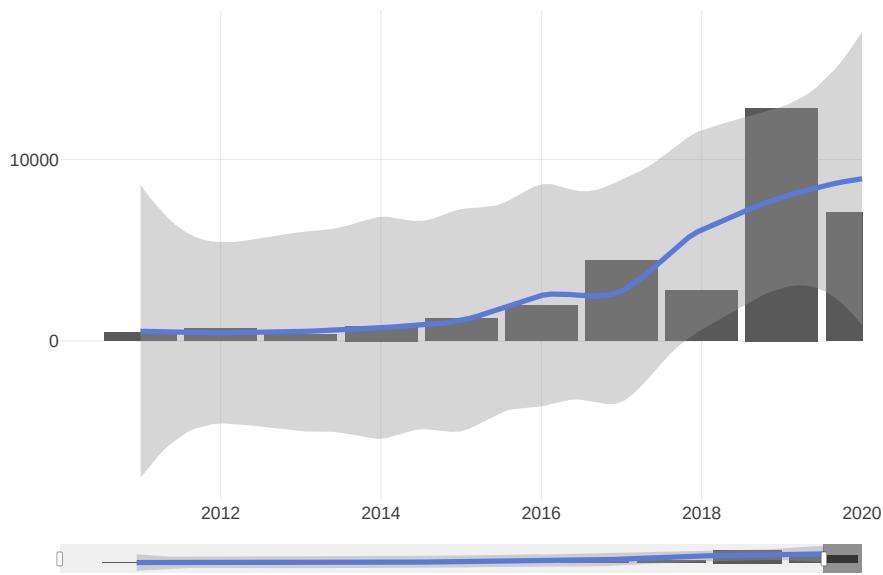
```

3.2 Looking at Change over Time

Following a basic cleaning, we can use the DOAB to examine broad trends and patterns in open monograph publishing. For example, consider this summary of open monograph publication volume over time:

[Hide](#)

```
library(plotly)
library(ggthemes)
time_plot <- { doab_df %>%
  group_by(`DT_ADDED_YR`) %>%
  summarize(total = n()) %>%
  ggplot() +
  aes(x = `DT_ADDED_YR` , y = `total`) +
  geom_bar(stat = "identity") +
  geom_smooth() +
  scale_color_fivethirtyeight() +
  scale_x_continuous( breaks = c(2010,2012,2014,2016,2018,2020)) +
  theme_fivethirtyeight() } %>% ggplotly()
time_plot %>%
rangeslider(start = 2010, end=2020, thickness=.05)
```



From this longitudinal overview, we can see that open monograph publishing is in its very early stages. Volume was quite small until five years ago, but has rapidly accelerated since then. However, progress is uneven, and after a record 2019 volume, there is a sharp downturn so far this year (noting that data for this year is not final) – possibly due to the pandemic.

3.3 Seeking Inclusion!

Since CREOS seeks to apply evidence to understand how disparate communities can participate in scholarship with minimal bias or barriers, it is of particular interest to understand the communities of authors that are currently included in open monograph publishing. The DOAB database includes additional information about each title, such as the year of original publication, names of authors, and subject fields (and we can add to that through linking to other sources through the ISBN) – however it contains no direct information about the characteristics of authors.

We can do better – making scholarship more inclusive requires making the characteristics of those participating visible: A more open & equitable scholarly knowledge ecosystem should support inclusion, self-descriptin, and information agency (Altman et al. 2018) Because no systematic public data on self-reported author characteristics exists, however, research on participation in scholarly publications must use bibliometric methods to impute gender from author names.(See, for example, Larivière et al. 2018 .) As an preliminary analysis, we apply a method that is commonly used in scientometric analysis and which is based on analysis of historical censuses (Blevins and Mullen 2015) to impute gender based on author names. We then use this imputation to explore the inclusion of works authored by men and women over time.³

[Hide](#)

```

library(gender)
library(genderdata)
library(humaniformat)
gender_meth <- "ipums"

doab_df %>% mutate(LS_NM_AUTHORS=str_split(`Authors`,";"))

# parse_names fails on empty strings, wrap it# gender can fail on genderize method
safe_first_name <- possibly(first_name, otherwise="")
safe_format_reverse <- possibly(format_reverse, otherwise="")
safe_gender <- possibly(gender, otherwise=list(gender=""))

doab_df %>% rowwise() %>% mutate(LS_NM_AUTHORS_R = list(safe_format_reverse (str_squish(`LS_NM_AUTHORS`)))))

doab_df %>% rowwise() %>% mutate(LS_NM_AUTHOR_FIRST=list(safe_first_name(`LS_NM_AUTHORS_R`)))

doab_df %>% ungroup() %>% rowwise() %>% mutate(LS_CAT_GENDERS = list(safe_gender(`LS_NM_AUTHOR_FIRST`,method=gender_meth)[["gender"]]))

doab_df %>% rowwise() %>% mutate(
  N_GENDER_MALE=sum(LS_CAT_GENDERS=="male",na.rm=TRUE),
  N_GENDER_FEMALE=sum(LS_CAT_GENDERS=="female",na.rm=TRUE),
)

```

[Hide](#)

```

library(rpivotTable)
doab_pivot_df <- doab_df %>% transmute (
  'Publisher' = `Publisher`,
  'Opened Year' = `DT_ADDED_YR`,
  'Any Female Authors' = `N_GENDER_FEMALE`>0,
  'Number of Female Authors' = `N_GENDER_FEMALE`
)
doab_pivot_df %>%
  rpivotTable(rows = "Any Female Authors", cols="Opened Year", vals = "Female Authors", aggregatorName = "Count as Fraction of Columns", rendererName = "Table Barchart"
)

```

Table Barchart ▾	Publisher ▾	Number of Female Authors ▾												
Count as Fraction of Columns ▾ ↪	Opened Year ▾													
Any Female Authors ▾		Opened Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Totals	
	false		60.4%	57.5%	58.1%	64.8%	62.4%	63.0%	69.8%	60.0%	68.7%	56.9%	64.4%	
	true		39.6%	42.5%	41.9%	35.2%	37.6%	37.0%	30.2%	40.0%	31.3%	43.1%	35.6%	
	Totals		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	

This preliminary estimate indicates that roughly thirty-six percent of open access monographs published in the last ten years have at least one female author. (This proportion remains roughly varies over time – but does not show a clear time trend.) As OA monographs are dominated by the humanities, where over fifty percent of US Ph.D. recipients (and over forty percent of faculty in most humanities disciplines) are women, this indicates a need to evaluate systemic bias of who is included in open monograph publishing.⁴

3.4 Follow the Money?

Business and economic models will need to evolve in order for monograph publishing to continue. The available data provides some hints (but only hints) on the economics of OA monograph production. The most comprehensive fully-open data is provided through the OAPC project and records book processing charges for the major consortial monograph purchasing initiatives.

We can use this data to look at fee-based revenue for presses participating in consortial open-monograph publishing arrangements. The most ‘profitable’ publishers are shown below:

[Hide](#)

```
library(lubridate)
## oapc cleaning
oapc_df <- mono_load_oapc()
oapc_df %<>%
  mutate(
    DT_ADDED_YR = year(parse_date_time(`period`, "y")),
    ID_ISBN_PRINT = lapply(`isbn_print`, isbnTools, meth="ean13"),
    ID_ISBN_MAIN = sapply(`isbn`, isbnTools, meth="ean13"),
    ID_DOI_ISBNA = lapply(`isbn`, isbnTools, meth="doi"),
  )
oapc_df %<>%
  rowwise()%>%
  mutate(LS_ID_ISBNS = list(
    setdiff(unique(c(`ID_ISBN_PRINT`, `ID_ISBN_MAIN`)), ""))
  ))
```

[Hide](#)

```

library(scales)
publisher_df <- oapc_df %>%
  group_by(publisher) %>%
  summarize(
    N_PUBS = n() ,
    TOTAL_REVENUE = sum(euro),
    AVG_CHARGE = TOTAL_REVENUE/N_PUBS
  ) %>%
  arrange(desc(TOTAL_REVENUE)) %>%
  mutate(publishers=str_trunc(publisher, 20))

library(crosstalk)

pub_key <- highlight_key(publisher_df)

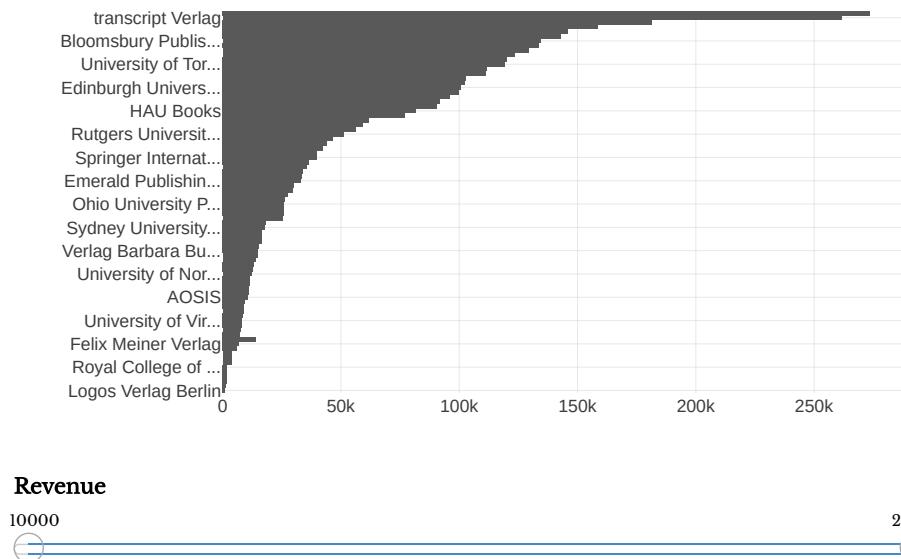
rev_plot <- { pub_key %>%
  ggplot(aes(x=reorder(publishers,TOTAL_REVENUE),y=TOTAL_REVENUE)) +
  geom_bar(stat="identity") +
  scale_color_fivethirtyeight() +
  scale_y_continuous(label=comma)+
  theme_fivethirtyeight() +
  labs(title = "Total Revenue (Euros) by Top Publishers", x = "Publisher", y = "Revenue") + coord_flip()
} %>%
ggplotly(dynamicTicks=TRUE) %>%
slice(1:25)

revSlider <- filter_slider("revenue", "Revenue",
  pub_key, "TOTAL_REVENUE", round=TRUE, dragRange=TRUE,min=10000, ticks=FALSE)

library(manipulateWidget)
combineWidgets(ncol=1,
  rowsize=c(9,2),
  rev_plot,
  revSlider
)

```

Total Revenue (Euros) by Top Publishers



Estimates of cost of producing monographs vary considerably, the most extensive study to date, estimated a range of average costs of approximately thirty to forty thousand dollars per title. (Maron et al. 2016) What does the OAPC data show?

[Hide](#)

```

library(plotly)
library(ggthemes)
library(manipulateWidget)

fees_plot_1 <-
{
  oapc_df %>% ungroup() %>%
    mutate(year = factor(DT_ADDED_YR, ordered = TRUE)) %>%
    ggplot(aes(x =year, y = euro)) + geom_violin() +
    scale_color_fivethirtyeight() +
    theme_fivethirtyeight() +
    labs(title = str_wrap("BPCs(€) by Year", width=15),
        x = "Charge (Euro)", y = "Year")
} %>% ggplotly()

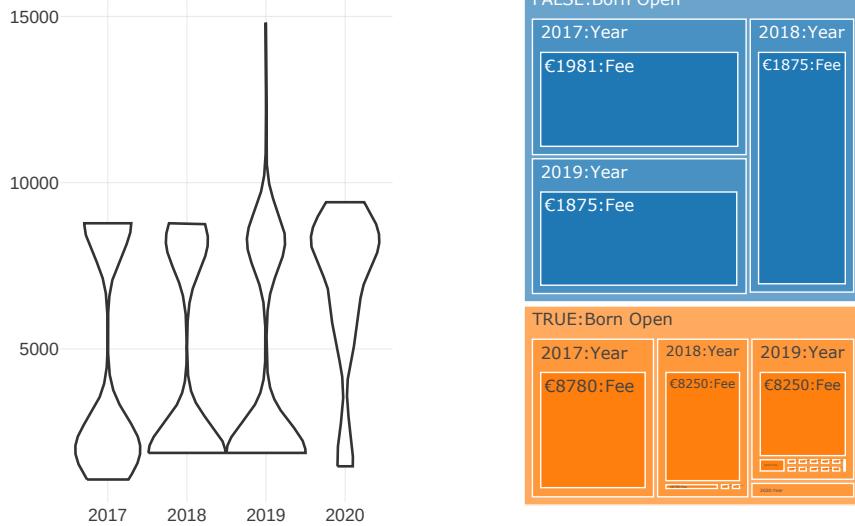
fees_plot_2 <-
  oapc_df %>% mutate(
    `Born Open` = !backlist_oa,
    `Year` = as.factor(DT_ADDED_YR),
    Fee = paste("€",euro,sep="")
  ) %>% xtabs( ~ `Born Open` + Year + Fee , data = .) %>% table_to_plotly_treemap()

# use combineWidgets, because subplot doesn't behave with treemaps, etc.
# see: https://github.com/ropensci/plotly/issues/655 and patchwork doesn't
# work with plotly

combineWidgets(fees_plot_1, fees_plot_2, ncol = 2)

```

BPCs(€) by Year



The most typical book charges in the data are approximately two thousand euros and ten thousand euros for converted-to-open and published-as-open monographs (respectively). These modes and the overall range is substantially under the range that would be expected from prior surveys.⁵

4 Future Puzzles ...

The exploration above raises a number of questions – under what conditions does the open availability of the monograph impact prices and sales? What are mediating factors – does the length or subject of the monograph mediate sales effects? What are potential mechanisms at play?

This exploration is limited by existing data. Each individual press has information on the sales, costs, and usage of the monographs they publish. If pooled, this data could potentially answer deeper questions about the economics and utility of academic monographs, and could guide a transition to open access models.

5 About this Document

This is a reproducible document. The most straightforward way to examine and modify the source is to clone the module using `git` and then load the project using `Rstudio`. The source is available here (<https://github.com/MIT-Informatics/monograph/blob/master/oamonoblog.Rmd>), and follows tidyverse style guidelines (using `styler` and `lintr` for conformance checking).

This analysis relies primarily on the `R` language, with `python` for the `ISBNtools` library. We make extensive use of the `Plotly` graphics package, and open R libraries (especially `tidyverse`, `gender`, `htmlwidgets`, and `crosstalk` and `Baker's R Makefiles`).

All references in this document are managed in `Zotero`. We use tidyverse style guidelines.

The authors describe contributions to this Essay using a standard taxonomy (see [allen2014]) Micah Altman provided the core formulation of the essay’s goals and aims, and led the writing, methodology, data curation, and visualization. Chris Bourg and Sue Kriegsman contributed to conceptualization and provided review. CREOS research assistant Shelley Choi provided assistance with preliminary data visualization and software implementation.

References

- Adema, Janneke, Graham Stone, and Chris Keene. n.d. “Changing Publishing Ecologies: A Landscape Study of New University Presses and Academic-Led Publishing: A Report to JISC,” 103. <http://repository.jisc.ac.uk/6666/1/Changing-publishing-ecologies-report.pdf> (<http://repository.jisc.ac.uk/6666/1/Changing-publishing-ecologies-report.pdf>).
- Altman, Micah, Chris Bourg, Philip Cohen, G Sayeed Choudhury, Charles Henry, Sue Kriegsman, Mary Minow, et al. 2018. “A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science.”
- Blevins, Cameron, and Lincoln Mullen. 2015. “Jane, John... Leslie? A Historical Method for Algorithmic Gender Prediction.” *DHQ: Digital Humanities Quarterly* 9 (3).
- Crossick, Geoffrey. 2016. “Monographs and Open Access.” *Insights the UKSG Journal* 29 (1): 14–19. <https://doi.org/10.1629/uksg.280> (<https://doi.org/10.1629/uksg.280>).
- Crow, Raym. n.d. “A Rational System for Funding Scholarly Monographs: A White Paper Prepared for the AAU-ARL Task Force on Scholarly Communication,” 34. <https://www.arl.org/wp-content/uploads/2012/11/aau-arl-white-paper-rational-system-for-funding-scholarly-monographs-2012.pdf> (<https://www.arl.org/wp-content/uploads/2012/11/aau-arl-white-paper-rational-system-for-funding-scholarly-monographs-2012.pdf>).
- Eve, Martin Paul. 2014. *Open Access and the Humanities: Contexts, Controversies and the Future*. Cambridge, United Kingdom: Cambridge University Press. <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/02BD7DB4A5172A864C432DBFD86E5FB4/978107097896AR.pdf> (<https://www.cambridge.org/core/services/aop-cambridge-core/content/view/02BD7DB4A5172A864C432DBFD86E5FB4/978107097896AR.pdf>).
- Grimme, Sara, Mike Taylor, Michael A. Elliott, Cathy Holland, Peter Potter, and Charles Watkinson. 2019. “The State of Open Monographs.” https://digitalscience.figshare.com/articles/The_State_of_Open_Monographs/8197625 (https://digitalscience.figshare.com/articles/The_State_of_Open_Monographs/8197625).
- Guédon, Jean-Claude. 2019. *Future of Scholarly Publishing and Scholarly Communication: Report of the Expert Group to the European Commission*. <https://doi.org/10.2777/836532> (<https://doi.org/10.2777/836532>).
- Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. 2013. “Bibliometrics: Global Gender Disparities in Science.” *Nature News* 504 (7479): 211.
- Maron, Nancy, Kimberly Schmelzinger, Christine Mulhern, and Daniel Rossman. 2016. “The Costs of Publishing Monographs: Toward a Transparent Methodology.” *The Journal of Electronic Publishing* 19 (1). <https://doi.org/10.3998/3336451.0019.103> (<https://doi.org/10.3998/3336451.0019.103>).
- Penier, Izabella, Eve, Martin Paul, and Grady, Tom. 2020. “COPIM Revenue Models for Open Access Monographs 2020.” <https://doi.org/10.5281/ZENODO.4011836> (<https://doi.org/10.5281/ZENODO.4011836>).
- Science Europe. n.d. “[Https://Www.ouvrirlascience.fr/Wp-Content/Uploads/2019/10/SE_on-Open-Access-to-Academic-Books_092019.pdf](https://Www.ouvrirlascience.fr/Wp-Content/Uploads/2019/10/SE_on-Open-Access-to-Academic-Books_092019.pdf).” https://www.ouvrirlascience.fr/wp-content/uploads/2019/10/SE_On-Open-Access-to-Academic-Books_092019.pdf (https://www.ouvrirlascience.fr/wp-content/uploads/2019/10/SE_On-Open-Access-to-Academic-Books_092019.pdf).
- Spence, Paul. 2018. “The Academic Book and Its Digital Dilemmas.” *Convergence: The International Journal of Research into New Media Technologies* 24 (5): 458–76. <https://doi.org/10.1177/1354856518772029> (<https://doi.org/10.1177/1354856518772029>).

1. The source for the document is available here (<https://github.com/MIT-Informatics/monograph/blob/master/oamonoblog.Rmd>). Since this blog takes the form of a fully replicable analysis, new versions can be generated as the data sources it relies on are updated. 

2. Monographs are typically uniquely identified through an ISBN, which is also a common choice when linking across databases. However, each ISBN is associated with specific formats (e.g. paper, hardcover, digital), so a single work published in multiple formats will have multiple ISBN’s. Further, the same ISBN may be expressed in multiple forms – so normalization is essential (`ISBNtools` is useful for this). Finally some databases will use DOI (digital object identifiers) or ASIN (Amazon standard identification number), instead of an ISBN. Generally the correspondence across identifiers must be resolved using an index: For DOI’s there is a programmatical mapping in theory to an ISBN13, but this often does not work in practice; and ASIN’s printed works generally match the ISBN number, but kindle editions (and related digital works) are assigned new ASIN’s. 

3. These imputations should be considered a very preliminary aggregate estimate, created for the purpose of promoting general discussion, potential issue spotting, and hypothesis generation. This method is intended for aggregate analysis and not for individual-level analysis – e.g. the assignment of an pronoun to an author. Further the reported imputation describes only point estimates, and does not reflect uncertainty from several sources: including omissions in the original data sources, heuristic name extraction, and uncertainty in name to gender assignment. Further, the analysis treats gender as a binary category, and thus will structurally omit non-binary gender categories. 
4. This is a formative, not summative analysis, and should be approached with caution. The gender imputation process contains many sources of unmodeled uncertainty; the analysis uses a US baseline, but the data does not support excluding non-US authors. Further this does not imply that bias in OA is worse than in scholarly publishing in general, since no baseline for gender inclusion in a comparable sample of non-open monographs has been established. The classification reported in the table is based on the IPUMS corpus. As a sensitivity check we evaluated using two other method: Use of historical Social Security Administration database yields a higher estimate of participation by at least one female author, but still lower than baseline expectation. Use of the popular ‘Kantrowitz’ method, which is based on a much smaller corpus – yields significantly lower estimates of female author participation. Notwithstanding – the range of estimates does not alter the overall substantive conclusions reported above. 
5. Note that the BPC charge does not necessarily reflect the entire cost of publication. However, the consortial initiatives included in the data above aim for the BPC to recover the costs of publication for born open materials. So the range of BPC charges should include the range of publication costs. 