Mapping the Free Ebook Supply Chain: Grant 11600118 (03/01/16 to 02/28/17)

Final Report to the Andrew W. Mellon Foundation, accepted 06/16/17

Charles Watkinson, Rebecca Welzenbach, Eric Hellman, Rupert Gatti, and Kristyn Sonnenberg

1.0 Summary of the Project and its Purpose

The number of open access academic ebooks is rising rapidly, with the *Directory of Open Access Books* catalog increasing from 2000 to almost 8000 listings over the last three years. As they become a growing (if still minor) feature in the marketplace, this project sought to better understand how free ebooks are discovered, acquired, and used by users. A particular focus was on learning whether free ebooks are succeeding in extending the reach of scholarly literature to users who might not otherwise encounter or be able to access it. Using a sample of books published by Open Book Publishers (OBP) and University of Michigan Press (UMP) (including some Open Humanities Press books), the researchers applied a range of quantitative and qualitative techniques: We compiled multiple sources of sales and usage data and conducted analyses on the dataset to try and find patterns; designed and deployed an open source survey tool to solicit responses from ebook users; conducted interviews with a sample of ebook users who had identified themselves through the surveys or on social media; and interviewed a range of vendors to understand how traditional supply chain partners are engaging (or not) with open access books. The team met regularly during the year, presented preliminary findings at several conferences, and is now working on a research report / journal article to share their findings more fully. The methodology and research results are interesting and the project members have also used their interactions through the project to raise recognition among publishers, libraries, and their supply chain partners of the challenges that new forms of digital publication (different business models, different formats) pose to the existing system. We believe that a concerted effort is still needed by all partners to ensure that the full potential of adopting open access approaches to supporting specialist scholarship is achieved.

2.0 Accomplishments

2.1 Usage data analysis reveals importance of marketing and need for library cataloging

Both OBP and UMP receive data from many different sources: Google Analytics information and web log data from their own platforms/websites; sales information for print and premium ebook versions; and (to varying degrees) usage data from retailers and aggregators. UMP is now also using Altmetric.com to track social media mentions and other indicators of online attention. Information from all of these sources was gathered for the sample books from the two publishers and supplied to the Free Ebook Foundation for exploration. The focus of analysis quickly moved to the web log data and Google Analytics since usage data from third parties turned out to be problematic for several reasons described below under “challenges.” The study of traffic and usage data was an exercise in characterization rather than statistics. It can be fairly said that every book had its own story and that aggregating the data was like mixing apples and oranges.

Usage data for both publishers can best be characterized as “spiky” both in the time domain and from book to book (see Appendix 1A). A few books get sustained usage but most titles exhibit extreme non-normal distributions (kurtosis over 1000) making statistical analysis problematic. Usage of books has long been known to tend towards Zipf-law distributions, with a small number of books getting a large fraction of the usage; the collections we studied were no exception. Large spikes can often be tied to particular events such as a mention in a news story or a social media mention. For most books the initial release spike is significant. This suggests that proactive marketing intervention by publishers is a necessary but not sufficient condition to drive usage.

Server logs and Google Analytics were used to study referrers, or sources of traffic/usage. The server logs revealed extensive robot traffic, both from a variety of general search engine spiders, but also from academically-oriented bots. One book (and one book only) was visited often by an unknown robot; more than half the usage of the book was from a single bot. Google Analytics was most useful for information about human referrals. As expected, the vast majority of traffic comes from Google; Twitter and Facebook are much smaller but still significant. What is striking is that very little use of either UMP or OBP books can be traced to library systems- catalogs, link servers, or discovery systems. There is a clear need to make sure that OA books are included more rigorously in library catalogs. The *Directory of Open Access Books* is the one significant source of traffic from the academic world, suggesting that it is playing an important role in discovery of these ebooks. The diagrams in Appendix 1.B show this pattern.

2.2 Survey responses suggest free ebooks do reach new types of reader

The Free Ebook Foundation created a survey tool designed by the team and links were embedded in several hundred OBP books and a smaller number of UMP titles. A copy of the survey questions is included in Appendix 2 and both code and survey example have been made available in a GitHub repository under an open license, as described in section 7.0. Each link was customized so that not only the title of the book that the respondent had accessed was reported but also the format (HTML, PDF) that they were reading in. The survey asked users how they had discovered the book, how they were using it, some demographic information, and whether they would be willing to be contacted for a follow up interview. By the end of May the team had received over 830 responses from OBP titles but only around 30 from Michigan due to fewer links in fewer books and challenges in making the survey link available from the Press's DLXS platform. A disproportionate response came from readers of a widely publicized ecology book from OBP. Several interesting patterns emerged from the survey data:

- ***OA books do seem to be reaching readers outside the academy:*** 50% of those who responded self-identified that they had acquired the book for "personal" use, with 44% "for my job," and 6% for "course use." Of those who shared their level of education 82% had graduate degrees and of those that shared their ages 62% were between 31 and 60, 29% were over 60, and 9% were 18-30. A third of the older respondents indicated that they were retired. Only 20% of respondents identified themselves as having established academic positions, with an additional 10% identifying employment within a library or university administration. In contrast, 34% of respondents identified non-university related employment, ranging across a diverse spectrum of occupations. Two thirds of respondents to the book attracting the greatest number of responses (*What Works in Conservation 2017*) identified themselves as having non-university employment. Some sample responses give a sense of who these "personal use" and "non-academic" readers are: "I'm a self-employed ecologist," "I'm a retired judge with a lot of interest in classics", "my personal use is in relation to family history." See Appendix 3 for a word cloud visualization of these responses.
- ***OA amplifies the diversity of different uses that ebooks already facilitate:*** Respondents were using the ebooks in a variety of ways; skimming, reading in detail, rarely reading the whole book, and often sharing. This last behavior was common (15% of the respondents planned to share the work) and reinforces the importance of social media and online communities in spreading the word about OA books. Readers want to use ebooks in flexible ways and with OA books they are not restricted from doing so by digital rights management. Some free-text responses indicate this behavior: "I've already shared the link with colleagues, and will discuss it with other people in my field, as appropriate," "I'll share any interesting highlights with others who are interested in the same thing," "I'll share it with other members of my NGO." A pie chart in Appendix 4 illustrates these patterns.
- ***Social media and "word of mouth" is extremely important in discovery:*** Releasing social media announcements and requesting that authors leverage their networks is well worthwhile: Over 35% of the respondents learned about a work from Twitter (15%), Facebook (9%), LinkedIn (8%), or Academia.edu (3%). Which one of these was the source of referrals varied by book, with LinkedIn

discussion groups (e.g., Higher Education Teaching Learning) and Facebook groups (e.g., British Ecologists Group) as important as direct sharing by authors. While coding the free-text responses was complex, email recommendations were probably the most common source of discovery. Traditional modes of discovery such as reviews in journals and citation tracking are mentioned but in a small percentage of cases. See Appendix 3 for a word cloud visualization.

2.3 Interviews with readers deepened our understanding of the survey results

Kristyn Sonnenberg interviewed 17 users of free ebooks in the sample who had either expressed their willingness to talk further when completing the survey or had discussed a book publicly on social media or in an online community. The small sample of interviewees was biased by which readers identified themselves and then which were willing to be interviewed but the aim was not to find a statistically valid sample but to provide more color to the picture that the survey results had started to reveal. Many of the themes emerging from the survey results were indeed reinforced through the interviews (the importance of social media as a discovery tool and readers embracing the flexibility of open access books to use them in a diversity of ways). However, some further themes of interest also emerged. These were not only about open-access vs. closed-access books but also about ebook vs. print book formats (a distinction that interviewees were more conscious of):

- ***Users of ebooks move to print to read the whole volume:*** Whether it was an open-access book or a purchased ebook, people were most likely to skim, read certain chapters, or read a book only to the point they needed to complete their work. As has been attested in other studies, multiple interviewees said they were more likely to read a paper book in its entirety than an ebook. As one interviewee reflected, “I think with printed books, I tend to have more of an in-depth, granular approach. With Kindle or ebooks, I tend to not want to look at them in the same depth, probably because screens aren’t design to be stared at for a long period of time.”
- ***There was strong interest in course use:*** Academic respondents were interested in using open-access ebooks as class texts, not only to save students money but also because it was easier to create syllabi with direct web links without worrying about availability or access controls. As one busy instructor said, “being able to throw links together from a bunch of different places and create a reading list for my class is good. I don’t have to put up with dealing with reserves or the bookstore.”
- **Users of OA ebooks do not tend to curate downloaded copies:** Several interviewees reflected on the chaotic organization of their electronic ebook libraries. “Why bother curating a copy when it is easy to find and download it again,” was the prevailing attitude. This behavior puts a burden on OA ebook publishers to make sure that copies are deposited in stable locations and preserved. Several interviewees mentioned that hyperlinking and other functionality was often lost after download or the OA versions they were reading were HTML only. As one interviewee said, ““I decided it wasn’t relevant to keep it or not keep it, since it was always available. What does it mean to keep a book if it’s open access?”

2.4 Interviews with infomediaries revealed how OA books challenge existing systems

Rebecca Welzenbach conducted a total of five extended interviews with six representatives from commercial players in the scholarly monograph information supply chain: JSTOR, Project Muse, Bowker (part of ProQuest), EBSCO, and Yankee Book Peddler/Gobi Library Solutions (now owned by EBSCO); a group we’ve referred to as “infomediaries.” The key aim of these conversations was to understand how these companies see their role in the OA information supply chain and what drives their decision-making around improving discovery of free ebooks. While the interviewees differed in how involved they are already in open access initiatives a few key challenges emerged across most or all of the conversations:

- ***Systems and workflows designed to lock down not open up books:*** These organizations operate complex systems, optimized for the sale of ebooks, with features such as complex digital rights management environment that make the accommodation of OA ebooks complex. While requests for changing these technologies may be simple, they are in fact complex and expensive and even OA advocates within these organizations must justify such investments with a solid business case (which is often lacking in this changing environment).
- ***The amount of OA book content is still small and will remain in the minority:*** Most of the interviewees work with a large volume of commercial content in addition to non-profit materials. OA (or potentially OA) monographs make up a very small slice of their business. Most anticipate that the OA share of the total market will likely always be small (5% was an estimate that came up several times), and investments in OA must be accordingly balanced with the requirements of the rest of their business.
- ***There are many unresolved questions about sustainability:*** Supply chain stakeholders want to see sustainable business models for OA and are not yet convinced that OA monograph publishing can be sustained for the long term at all without "unstable" interventions such as philanthropy or grants. A lack of knowledge about what is happening outside North America, and the more centralized funding models evolving in Europe, was notable. Preparing books for deposit in platforms, including cataloging, incurs costs and it was unclear whether publishers or libraries would be willing to fund these fees.
- ***Variance in definitions/practices around OA make systematization hard:*** Inconsistency in how publishers define and implement open access is a real problem for supply chain stakeholders: expectations and requirements vary a lot from publisher to publisher, and designing systems to support OA content is difficult when there are many different potential scenarios to support. For example, one interviewee cited the variation in attitudes among publishers as to whether the PDF download of a book should be free.
- ***There is a lack of consistency in metadata:*** Interviewees noted that that current metadata standards in use (ONIX, MARC, etc.) do not sufficiently support publishers in sharing information about whether a book is open access or not and, if so, on what terms. There was concern about the use of different metadata standards among publishers, libraries, and infomediaries.

Despite these challenges, interviewees express a strong desire to demonstrate or reinforce value to libraries and if library clients support the expansion of OA book content they will be responsive. They recognize their role(s) as a trusted library content provider/discovery platform, and acknowledge that in order to maintain that position, they will need to continue to select content carefully, include all the best material on their platforms irrespective of business model, and add value in the form of services and tools. They are eager to do this in a way that is efficient and responds to customer demand. During the period of the grant, and perhaps partly due to our advocacy work, we saw several vendors make changes to their programs to support OA books better and continue to consult with them to improve such support.

3.0 Setbacks and Challenges

3.1 We encountered obstacles in obtaining and mapping data from the different sources

OBP and UMP distribute through different vendors, only receive usage data from some and not others, and receive usage data that is not comparable (for example, some platforms present only chapter downloads while others present whole book downloads). The differences in the ways that delivery platforms and websites are organized makes it difficult to tell a data-driven story of the impact of OA approaches. Google Analytics and web log data are more comparable but the privacy policy that University of Michigan Library has put in place, and the stripping of identifying information, impeded some of the analysis that we wanted to do.

3.2 There were obstacles to deploying the survey widget

OBP deployed the survey in many more places and with more books than UMP, taking advantage of the greater focus they have on delivering books through their own platform. Some third party platforms disable URLs (for example, Google Books and BiblioBoard) which meant that readers clicking on the links did not get a response and were not able to get to the survey. It was difficult for UMP to retrospectively insert links in OA books after files had already been distributed through Knowledge Unlatched and then harvested and redeposited by other platforms. This is again a reminder that the publisher rapidly loses control of where openly-licensed content is hosted which would made any other form of content correction hard to manage.

3.3 Analyzing survey data was time-consuming because of uncertainty what the responses would be.

We received rich responses but because we didn't know going into the research what they would be it was difficult to use controlled vocabularies or other such tools that would have made automated analysis easy. Visualizing and understanding the responses in may be most effectively accomplished through a device such as a Word Cloud as shown in Appendix 3. Now we have a sense of the types of answers that were most common it will be easier to deploy drop-down menus to replace some free text fields in a future iteration of the survey.

3.4 Commercial confidentiality makes interviewing commercial competitors challenging

We anticipated that our main challenge would be to devise questions that would garner insights into how companies are thinking about open access, without causing interviewees to feel we were prying into proprietary information. We found in fact that interviewees were quite candid, often offering more specific information than we asked for. But we are still limited in our presentation of the data to provide generic overviews.

3.5 Finding a diverse range of users to interview was difficult

Outreach to potential interview subjects took a great deal more time than we anticipated. We found some interviewees by way of the survey when users shared their contact details, but others we needed to seek out based on their public book reviews on GoodReads, their social media posts about the book, etc. In the end we sent out 70 invitations (each recipient carefully identified and selected) to get to 17 interviews. Due to language and time-zone constraints, our interviews were largely limited to North America and English-speaking subjects.

4.0 Significant Staff Changes

The interviews with ebook users were originally going to be done by Audrey Evans. She got another job offer soon after the grant had started. Dr. Kentaro Toyama in the U-M School of Information helped us to identify graduate student Kristyn Sonnenberg who kindly agreed to take over Audrey's role on the grant.

5.0 Plans and Goals for the Period Subsequent to the Grant

5.1 Creating a more robust and widely deployed survey tool

The number and richness of responses to the survey we imbedded in OBP and UMP books surpassed our expectations. Users seemed much more ready than we had expected to share information in gratitude for receiving a free book. While the technology generally worked well there are some improvements that we would like to make in the survey tool so it could be more broadly used by the community. These include more programmatically generating the links for imbedding in books and expressing them in short forms that could be used even if hyperlinking functionality in some ebook readers is disabled. We would also be interested in exposing the survey creation user interface to potential users and adding more summary and reporting features.

The European Union-funded Hirmeos project (http://operas.hypotheses.org/hirmeos), with which Rupert Gatti is involved, is interested in creating tools and infrastructure to advance OA book publishing and may be interested in further development of the tool.

5.2 Structuring survey to allow automated analysis / linking web logs and survey results

While we believe that the deliberately-relaxed tone of the survey, its length, and the questions asked were all conducive to getting a high level of responses, the analysis would be made easier through the use of more controlled vocabularies, for example a list of professions. If we obtained funding to strengthen the survey tool, we would also restructure the survey and get broader input on the sorts of questions other OA publishers would like to ask. A customizable tool that would let publishers create their own version and deploy it in scalable ways that worked for their particular systems would be the end goal. A tantalizing question we would like to pursue is the degree to which we can combine information about what users say (from the survey) and what they do (from log data) by associating IP addresses. The technical challenges can be overcome but there are also some pressing legal and policy questions about protecting user privacy. We would be interested in trying to make these associations in order to learn more about reader behavior (e.g., do retired people read the whole book while students just look at one page?) and improve the reader experience. We are also interested in geocoding responses.

5.3 Larger scale program of interviews of open ebook users

While we debated whether to formally code the interviews conducted with ebook users, we concluded that the sample size was not large enough to do so. We would be interested in expanding the size of the study to also include a more diverse and statistically robust sample, including users in developing countries where the adoption of OA resources is of particular interest to open access advocates. It may be possible to connect this project with the further development of the survey tool described above since a larger number of survey responses would increase the population of willing interviewees.

5.4 Aggregation, analysis, communication of usage data from more platforms

Eric Hellman is documenting the techniques he used to conduct the usage data analysis so this information can be useful to other projects. There are already several initiatives working on the analysis of usage data from open access books, most notably Knowledge Unlatched Research. We see a particular opportunity to collaborate with KU Research around the challenge of aggregating usage data from multiple sources, analyzing the data to create insights, and then communicating information to important stakeholders including authors and funders (increasingly likely to be the parent institutions of the authors). The two main components of an aggregation of data (a “data warehouse”) and then tools to analyze and communicate custom views of the data (a “data dashboard”). While publishers are working on creating tools to do this work independently, we would be interested in seeking grant funding in collaboration with initiatives such as Project Meerkat and KU Research to create a warehouse and dashboard tool based on open principles.

5.5 Advocacy with infomediaries

One of the most useful aspects of the interviews we conducted with infomediaries is the opportunity these created to advocate for more visibility of open access books within their systems. We are continuing to consult with EBSCO, ProQuest, JSTOR, and Project Muse on the best way to incorporate open access books into their products. We would like to expand the conversations to include UK and European ebook vendors (e.g., Casalini Libri, Harrasowitz, Gardners, Bertrams) who may be more alert to the growth of OA and to other, less visible, participants in the supply chain -- such as preservation services such as Portico, or review destinations such as CHOICE. As we investigated information flows we found that free ebooks could be excluded from being

preserved or reviewed because important players depended for metadata and files on supply chain vendors who made their money from book sales. A particular issue we uncovered was the lack of a field in MARC that would indicate whether a particular book was OA or not. OA availability information is shared by an increasing number of publishers following the OA Monographs in ONIX standard, but the mapping between ONIX and MARC is notoriously problematic. We are working with OCLC and the technical services department at University of Michigan Library to improve the way in which open access is flagged in bibliographic records.

6.0 Publications, news articles, or other materials related to the grant

We are currently preparing a fuller report for publication in the *Journal of Electronic Publishing* which will include further details of our work. Preliminary information was shared in presentations at the *Charleston Library Conference* ("Mapping the Free Ebook Supply Chain" November 3, 2016) and at the *Coalition for Networked Information Fall 2016 meeting* ("Building the Better Ebook and Beyond" December 13, 2016).

The article for *JEP* has been tentatively accepted by the editor, Maria Bonn, pending final editorial review. The final submission is due by end of July 2017 and *JEP* estimates that it will be published by end of August 2017 based on current commitments. This article and the white paper described in the grant proposal are the same thing. We felt *JEP* was the best place to share results because the journal is open access and highly read within the scholarly communications community.

7.0 Intellectual Property

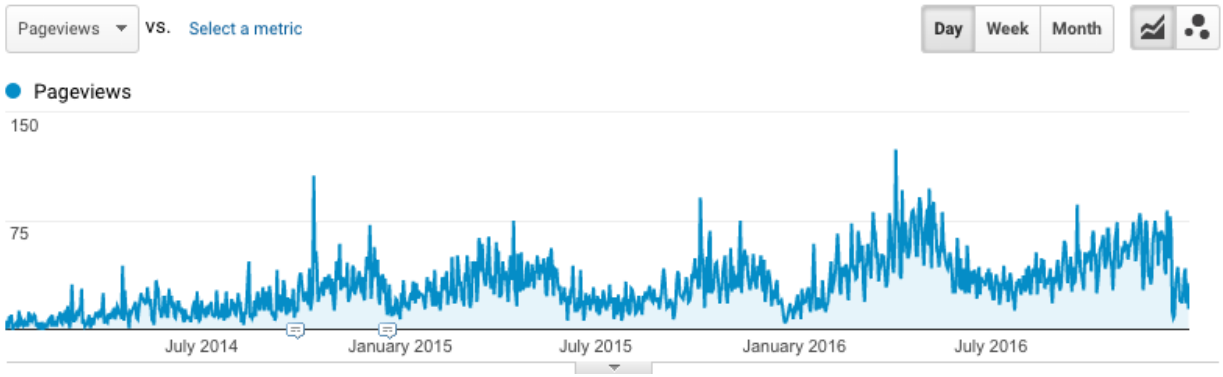
As described above, the Free Ebook Foundation created a survey tool designed by the team and links were embedded in several hundred OBP books and a smaller number of UMP titles. The survey tool is a "Django" app derived from an app originally developed by "Seantis" and modified by "Eldest Daughter". We updated it to work with the current version of Django, and added features to allow each response to be tied to a book catalog. The code for the survey tool is available on github at https://github.com/EbookFoundation/fef-questionnaire and is carries a "Modified BSD" License. The survey questions have been loaded into an "example" application "fixture" that is also in the GitHub repository. JSTOR has already expressed interest in reusing code and survey questions, adapting and developing them for its own use.

Appendix 1: Data from Google Analytics

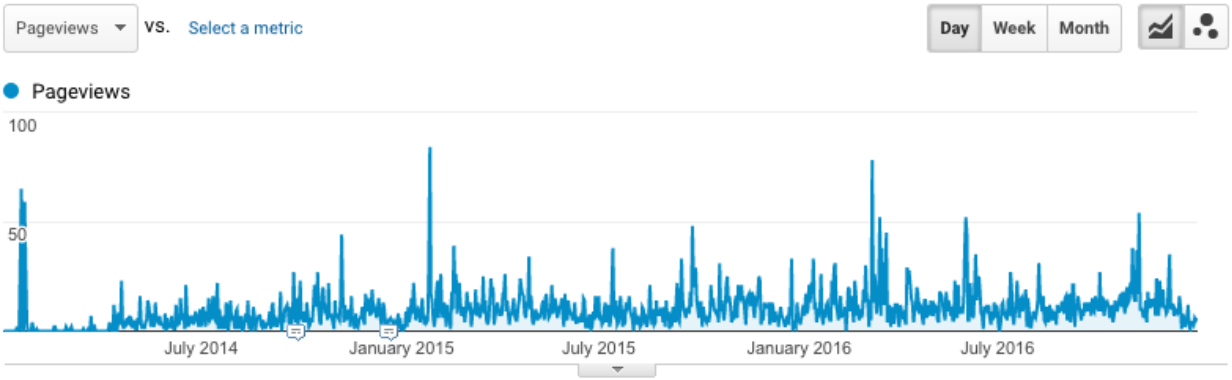
A. Daily Traffic

Traffic patterns are shown for two titles from the Open Humanities Press imprint, distributed by University of Michigan Press

Three years of daily page views for “Bare Life” show seasonal traffic. The bulk of the usage is for a single chapter, possibly being used as course material. Only 3 or 4 books of the 137 books studies were as smooth as this.

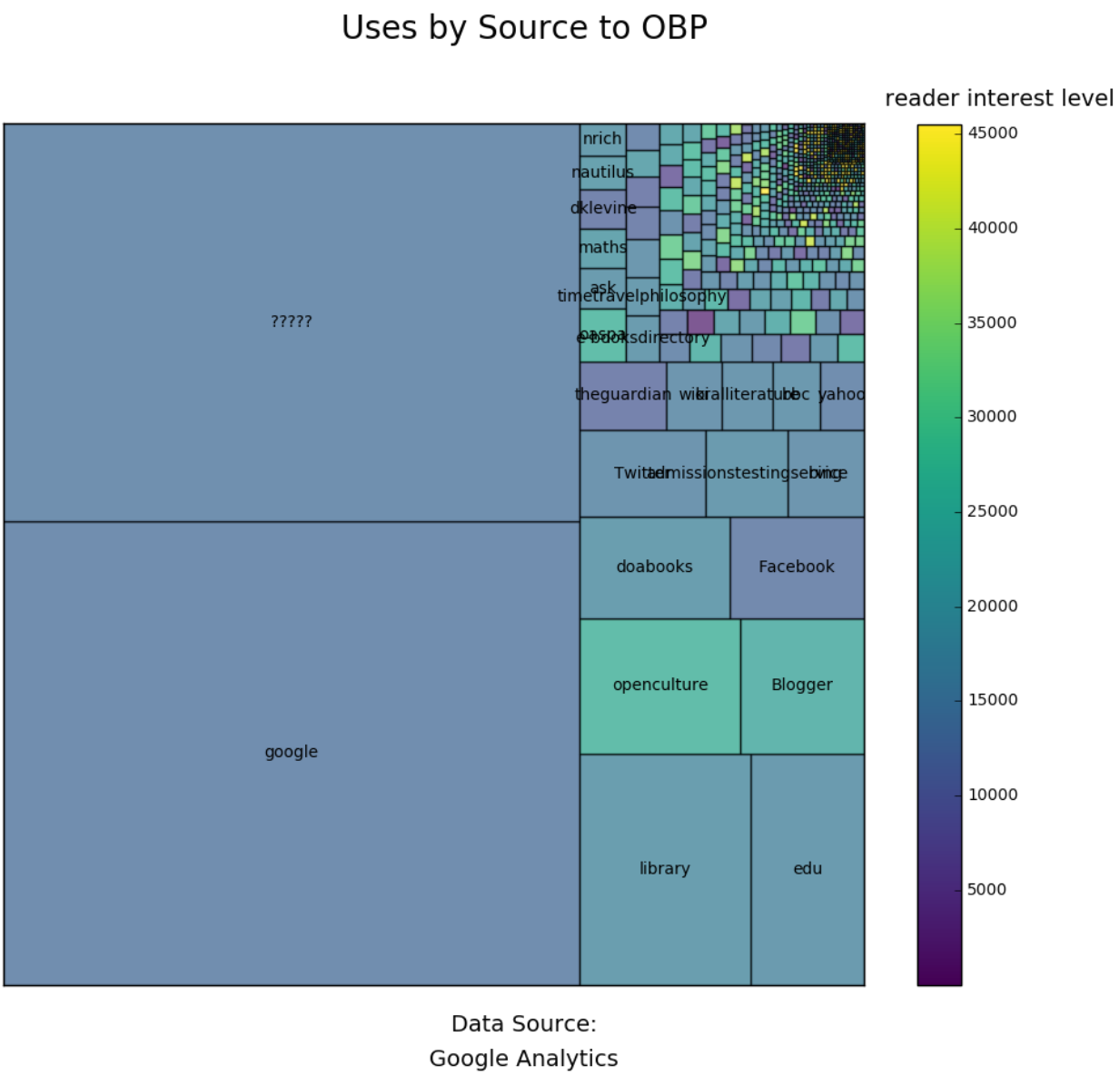


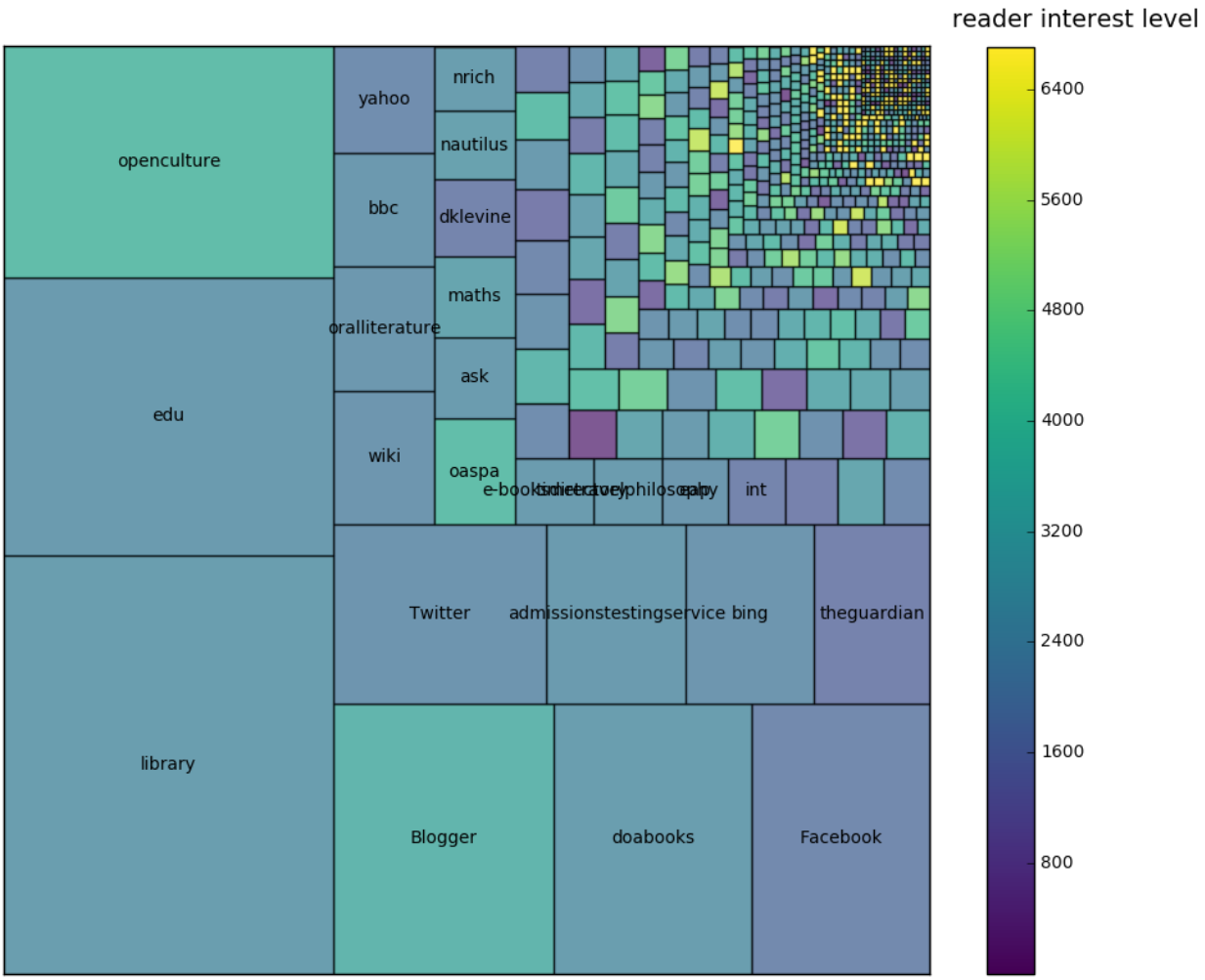
Three years of daily pageviews for “Architecture in the Anthropocene” show spiky traffic, including a release spike in January 2014. This spiky usage is typical of most of the books studied.



B. Referrer Information

The diagrams below show traffic to the Open Book Publishers website, the first including all visits and the second with unidentified referrer ("?????") traffic and traffic from Google removed. Similar results were obtained for University of Michigan Press websites. The color of each box indicates the reader interest level, as measured by bounce rate.




Uses by Source to OBP
reader interest level
6400
5600
4800
4000
3200
2400
1600
800
openculture
edu
library
yahoo
bbc
oralliterature
wiki
nrich
nautilus
dklevine
maths
ask
oaspa
int
Twitter
admissionstestingservice
bing
theguardian
Blogger
doabooks
Facebook
Data Source:
Google Analytics

Appendix 2: Survey Questions and Results

Welcome, reader of _____! And thanks for visiting Unglue.it to complete this survey, part of a research project to understand how open access ebooks are discovered and how readers use them. For more information, please see the project description.

As Open Access publishers, _____ are truly committed to making academic research broadly accessible - so we want to understand how people like you are actually accessing and using our Open Access titles.

We have a bunch of questions for you (well - only 10 actually) about how you found this book and what you're going to do with it. Please tell us the things you think are interesting or relevant. We really want to know!

[Privacy policy: There are no marketing traps, we're not going to spy on you or swamp you with emails afterwards, or tell our "friends" about you - we're just going to store your answers to create a database of usage examples that can be used to understand what Open Access publishing enables. A report of the results will be made available in July 2017]

About the Book:

1. How did you find out about this book in the first place? For example: Was it from a Google search? Following a wikipedia link? A tweet? Referenced in another book? A late night session with a friend? - or in some other way? {Free Text Box}

2. How did you get hold of this particular copy? For example: Did you download it from the publisher's website? Amazon or another retailer? Find it on academia.edu? Or somewhere like aaaaarg? Get it from a friend? Your library? {Free Text Box}

3. Why are you interested in this book? {Multiple choice}
- For personal use - I'm interested in the topic
- For my job - it relates to what I do
- I need to read it for a course
- Other – If other, tell us more...

4. Are you aware that this title is available in multiple different digital and printed formats? {Yes/No} If Yes - is there any particular reason why you are using this version rather than one of the others? {Free Text Box}

5. What are you going to do with it now you have it? {Multiple choice}
- Save it, in case I need to use it in the future
- Skim through it and see if it's at all interesting
- There's only really a section/chapter I'm interested in - I'll probably just read that
- The whole book looks fascinating - I'm going to read it all!
- I'm going to adapt it and use it (or, at least, parts of it) for another purpose (eg a student coursepack, lecture/briefing notes …)
- Share it with my friends
- Print it out
- I'm creating/collating a (online) library
- Something else entirely …. Please tell us in more detail:

Now About You...

And now, four questions about you as well ...

6. Where do you live? {Multiple choice}
- Canada/USA
- Central America/ Caribbean
- South America
- Europe
- Middle East
- Africa
- India
- China
- Other Asia
- Oceania
- Another Planet
- Other...

7. What do you do for a living? {Free Text Box}

8. How old are you? {Multiple Choice}
- under 18
- 18-30
- 31-60
- over 60
- decline to say

9. When did you finish your formal education? {Multiple Choice}
- I haven't - I'm still a student
- At primary/elementary school
- At high school/secondary school
- After trade qualifications
- At College/Undergraduate Degree
- At Grad School/post-graduate university
- Other...

10. Is there anything else you would like to tell us or think we should know about how you found or are using the ebook? or about yourself? {Free Text Box}

Follow-up

We would really like to be able to follow up with some of the respondents to this questionnaire to ask them a few more questions - particularly if you've told us something really interesting in a comment (for example). [There will also be a little reward (a free book no less!) for those of you we do contact in this way.]

Thanks so much for your time and efforts answering these questions for us - we love you for it!

We hope you enjoy the book. If you're willing, then please leave us an email address where we could make contact with you (information which we won't share or make public). {Free Text Box}

Appendix 3: Word Cloud Visualizations

This shows two sample Word Cloud visualization of responses to questions in the survey. In the first example, the preponderance of word of mouth and social media recommendations as a means of discovery emerges clearly. In the second example, the academic and retired bias in respondents is revealed.

Appendix 4: Pie Chart showing variety of uses

This pie chart shows how respondents to the survey plan a variety of uses. Most are familiar from studies of gated ebook use but the emphasis on sharing is notable since this is a desired behavior facilitated by OA availability.