Mapping the Free Ebook Supply Chain

Proposal Narrative

Summary Description

The University of Michigan requests $27,656 to conduct an evidence-based study of how free ebooks are discovered and used. The study would be conducted over the course of a year, between March 1, 2016 and February 28, 2017. Working with a diverse sample of around 120 free ebooks published by Open Book Publishers (OBP) and University of Michigan Press (UMP) along with Michigan-distributed titles published by Open Humanities Press (OHP) we will engage in retrospective and prospective analysis using both quantitative and qualitative metrics.

In the retrospective analysis, we will examine every available quantitative source of usage data for the sample books that have been available for at least six months and will conduct semi-structured interviews with some key stakeholders to gather data about how the sample books were discovered and are being used. To facilitate prospective analysis, UMP and OBP will implement a listening/assessment exercise by inserting a link to a web-based usage survey in the front matter of books that they publish starting in April 2016.

Why is this project important? A number of initiatives focused in North America, Europe, and Australia are proposing to convert the system through which scholarly books are currently funded from a consumer- to producer-pays model which will allow free ebook versions to be made available. One driver for this change is the conjecture that releasing an ebook under an open access business model expands its readership compared to "toll-access" or "pay-to-purchase" ebooks. This is a reasonable hypothesis but we actually know very little about how free ebooks are discovered and lack best practices and tools to measure their impact. This is the challenge that the proposed project aims to engage with.

While the "open web" is known to be the most effective information distribution infrastructure ever built, a conventional supply chain relying on intermediaries taking proportions of retail price continues to dominate the book market. Starting from a study of current and prospective users of free ebooks we aim to understand whether the conventional supply chain remains important or whether other mechanisms of discovery are sufficient. We will also interview representatives of supply chain companies (including jobbers, retailers, and aggregators) to
understand how they view free ebooks and whether they see a role for their organizations in discovery and delivery.

The outcomes of the study will be (a) a white paper sharing its findings out of which we will develop an article that we will submit for publication by a leading journal such as the *Journal of Electronic Publishing*. The white paper and article will include recommendations about best practices for ensuring discovery of free ebooks and meaningfully measuring their impact; (b) an open-source web survey application that publishers can use to capture qualitative information about ebooks usage.

*Background*

How do readers discover ebooks that are free? How do they obtain and use them? Who are these readers? When books and ebooks carry a purchase price, a conventional supply chain connects publishers with readers and tracks the purchases so that revenue can be returned to the publisher. Participants in the supply chain work to enhance discovery and identify markets, and they are compensated for their efforts with a percentage of the revenue. Free ebooks are often denied entry into this supply chain and tracking the users and the usage is much more difficult.

It is conjectured that releasing an ebook under an open access business model expands its readership compared to "toll-access" or "pay-to-purchase" ebooks. This is certainly a central tenet of current initiatives to expand the availability of open access monographs. In articulating the joint proposal by AAU and ARL for a system based on institutionally funded faculty book subventions, Raym Crow identifies the desire to "increase the visibility, discovery, and measurable use of scholarly research by providing open-access dissemination, thus helping to accelerate changes in tenure publication policies and practices."[1] Meanwhile, in his HEFCE report on monographs and open access, Geoffrey Crossick "recognises that open access has a great deal to offer, not least in terms of increasing the reach and impact of research publications and improving both the extent and the character of scholarly communication within the academy."[2]

That freely accessible ebooks expands readership is a reasonable conjecture because the "open web" is known to be the most effective information distribution infrastructure ever built. However, the conventional supply chain continues to dominate the book and ebook markets and it has hundreds of years of infrastructure development behind it. For example, it is quite difficult to get an open access ebook into the Amazon Kindle ecosystem, and there are significant barriers in the Apple iBookstore as well. Most libraries use aggregators such as EBSCO, ProQuest, JSTOR, Project Muse, and Overdrive to manage ebook provision, each of
which presents its own sort of barrier to open access ebooks. Because selection and acquisition processes are largely outsourced to commercial “jobbers” such as Yankee Book Peddler and Coutts, libraries tend to overlook open access ebooks. The preservation and maintenance of open access ebooks are also disorganized, since standard industry solutions like Portico mostly rely on batch relationships with publishers and aggregators financially based around “toll-access” publishing models to receive deposits.

Open access ebook publishers have a limited view of ebook usage. As measured by downloads from the publisher’s website, usage typically outstrips book sales by impressive multiples. But close examination of download logs indicates that many downloads involve automated downloads or repeated use by individual users. Open access licensing also allows the books to be downloaded from third-party websites, institutional repositories, and aggregation sites, resulting in much invisible usage activity. As researchers from Knowledge Unlatched note, “the increased discoverability that results from books being hosted on diverse platforms can only be beneficial, however combining usage data from multiple sources creates data blending challenges not only for books in the KU collection but for all OA books.” Mobile ebook formats result in further difficulties in interpreting usage, as reader apps often use machine interfaces rather than web browsers to access ebooks.

It’s easy to identify some important participants in the free ebook supply chain. Certainly the publishers offering the ebooks for download directly from their websites are vital links. Certainly Google’s search engine is an important discovery channel for these ebooks. Social media, both popular networks such as Facebook and Twitter are significant too, but it’s hard to say how much. But beyond these, we don’t know for sure what the free ebook supply chain comprises.

Open access publishers, authors, and stakeholder institutions and funders need to have a better understanding of free ebook usage in order to know how to increase it. The success of a book has tended to be measured by authors and publishers based substantially on the sales it achieves. But in an Open Access environment sales data is much less representative and the new currency of impact must be usage data. The ability for publishers to aggregate and communicate meaningful usage data is essential to the sustainability of an open access business model: Authors need to be able to demonstrate the impact of their books, institutions need to be convinced of the quality of the scholars and presses they support, and funders (increasingly including institutions) need to be able to measure return on investment.

---

1 http://www.knowledgeunlatched.org/2015/12/ku-research-update/
This study proposes to take first steps toward achieving this understanding of usage, exploring who users are, how they discover books, and how their activities can be tracked and communicated.

**Schedule of Major Activities**

This effort will go in two directions, one retrospective and the other prospective.

The retrospective effort will characterize the distribution of a total of around 120 books that have been published for at least six months (allowing some usage data to have accreted). Twelve of these books will be selected for a "high resolution" study.

For each of the larger set of books, we will gather and collate available sales/usage/download statistics and attempt to identify points of discovery and distribution. Quantitative data gathered will include web analytics, download logs, sales data (for premium electronic and print editions), aggregator data (where obtainable), library holdings records, social media references, Google AdWords and webmaster data, reviews and citations, records of alternate hosting in repositories, etc., and bookstore listings.

Titles to be studied will include titles that have been available in multiple formats; at least print, pdf, and epub, and some that are available as web versions. We will include titles with differing open access formats and licensing statuses (e.g., different Creative Commons licenses). The number of books to be studied provides some diversity in age, subject matter, target audience, and format. Quantitative measures of usage such as sales and download counts will be analyzed using multifactor component analysis to assess the relative influence of these factors. The analysis of download logs will supplement counting statistics for the entire corpus with in-depth analysis for the high resolution set.

In terms of the data from traditional aggregators, there is a large overlap between the channels in which OBP titles appear and those which UMP and OHP titles are in, although each publisher has emphasized slightly different relationships. At the minimum all aggregators provide ISBN-by-ISBN sales/usage data on a regular schedule so no additional negotiation with or permission from the channel partners would be needed. Because only EBSCO and Project Muse currently provide information about the end user without further request (and even then it is the library who purchased a book rather than the reader who used it) the limit of analysis for traditional aggregators will be sales by ISBN. Points of discovery (vs. delivery) will be investigated through other sources of data, such as web logs and Google Analytics.
The rough plan for the data collation and analysis by the Free Ebook Foundation Team is as follows:

- Data collation (14 days total). Per-book data from 120 books will be collated into one big table.
  - Initial title list cleanup (1 day)
  - Additional classification (1 day)
  - Log file processing (1 day per log file source x 3 sources) (3 days)
  - Google Analytics data extraction (including HathiTrust) (2 days)
  - Aggregator data extraction (1-2 hours per aggregator) (2 days total)
    - Google Books (GooglePlay)
    - OpenEditions
    - Wikimedia
    - Internet Archive
    - WorldReader
    - LightningSource
    - EBSCO, including NetLibrary
    - Proquest, including Ebrary, EBL, and MyiLibrary
    - StreetLib (aggregates sales from Apple iBookstore, Kobo etc)
    - Amazon
    - Project Muse/UPCC
    - OApeng/DOAB
    - Feedbooks, Manybooks
  - Social media references (2 days) May include
    - Mendeley
    - Altmetric.com
    - Academia.edu
    - ResearchGate
    - LibraryThing
    - Twitter
  - Google AdWords and webmaster data (1 days)
  - Reviews and citations (reviews from Google Books, citations where available from Google Scholar and the sites it indexes (SSRN, etc.) (1 day)
  - Search for records of alternate hosting in repositories via WorldCat (0.5 days)
  - Other bookstore listings (0.5 day)

- Analysis Phase (8 days total).
  - Multi-component analysis using R (2 days)
  - High resolution study; 2 hours each for 12 books (3 days)
  - Summarize results (3 days)
For the high resolution set, we will attempt to connect downloads with usage by individuals and by machines by resolving IP addresses and studying usage patterns. In all cases, website privacy and security policies will be adhered to, and no collected data will be exposed except in aggregated, anonymized forms.

To supplement the quantitative part of this study, we will conduct two sets of interviews.

The first set of interviews will focus on users and potential users of the books under study. The focus will be on modes of discovery, usage and acquisition of the studied books. The interviewees will include: authors of the studied books, researchers working in the fields covered by the studied books, subject specialist librarians in the fields covered by the studied books, people who cite a studied book or similar books on social media, people identified by their response to the in-book assessment link, teachers/lecturers who have assigned works to reading lists, and representatives of developing country readers/users.

A second set of interviews will look at the evolution of the supply chain. We will ask representatives of supply chain organizations about their visions for how the supply chain for free ebooks should work in the future in an “ideal world” and what role they would see for their organizations in that world. We believe that phrasing questions in such a hypothetical/vision way will avoid sensitivities about revealing competitive intelligence. Participants will be selected from the traditional supply chain and from entities we identify from data analysis. In selecting interviewees we will define “supply chain” widely to include, for example, review venues like CHOICE and preservation solutions like Portico and HathiTrust.

The prospective effort will create listening/assessment tools that work from inside newly published books as well as from a website. These will consist of book-specific links or buttons that can be added to front matter which request that readers of the book support the book’s publisher by clicking the link and filling out a brief usage survey. The survey will ask about channels of discovery and distribution, and basic, non-identifying demographic information.

To process user responses, we will create a survey application based on off-the-shelf open source software modules which support configurable surveys. Survey responses will be integrated with a book catalog so that responses can be analyzed for each book or group of books. For the current study, we will initially use the Unlue.it book catalog and website to host the survey, but the survey application can be fully integrated into any book catalog and website run using the “Django” application framework, and can be run alongside other web frameworks on a wide variety of servers. The data format will depend on the particular survey module used, and will be exportable and sharable via a REST API.
The usage survey will start with some demographic profiling (e.g., what is your occupation? would you describe yourself as an academic researcher/teacher or not? is this title of interest to you in your “professional” or “personal” capacity?) followed by questions about how interviews used the books such as: How did you find out about this title? How did you discover the URL for the title? How did you use this ebook, e.g., for reference, background, or citation? Did you read it, just skim it, or focus mostly on one section? This volume is available in multiple formats - why did you select to use this edition of the title? Did you share the item/URL with anyone else? Did you download the item for future use, or make note of the item in another system some other way?

Many points of discovery and points of distribution will be easily discoverable, others with be completely hidden. When UMP or OBP are the point of distribution, we can look at referrer headers in log files to (partly) trace back to the point of discovery. This is why the survey tool is so important to this work; we can ask about points of discovery (such as personal communications) or distribution (private archives or libraries) that are otherwise completely hidden. We can’t avoid the lamppost effect completely, but by looking out from various directions, we should be able to estimate the area lit by the lamppost.

**Participants**

**Charles Watkinson** is Associate University Librarian for Publishing at the University of Michigan and Director of the University of Michigan Press. Since 2006 UMP has published almost 50 open access books in its Digital Culture Books imprint as well as being a leading participant in the Knowledge Unlatched program. Watkinson is jointly leading a project entitled “Building a Hosted Platform for Managing Monographic Source Materials and Born Digital Publications through Library/Press Collaboration” generously funded by the Mellon Foundation to build a software platform optimized for multimodal book publication with robust usage aggregation and reporting an important component. Watkinson will act as co-PI. He will lead the writing of the white paper and recommendations and manage reporting and tracking on the grant.

**Eric Hellman** is a technologist and entrepreneur who leads the Free Ebook Foundation, which “envisions a world where ebooks will be funded, distributed and maintained for the benefit of all, by coordinating the efforts and resources of many.” He is also founder of Unglue.it by Gluejar which provides a platform for CC-licensed ebooks as well as providing innovative crowdfunding technologies to help publishers make their books open access. After 10 years at Bell Labs in physics research, Hellman became interested in technologies surrounding e-journals and libraries. His first business, Openly Informatics, developed OpenURL linking software and knowledge bases, and was acquired by OCLC in 1996. At OCLC, he led the effort to productize and expand the xISBN service, and began the development of OCLC’s Electronic Resource
Management offerings. He started Gluejar after leaving OCLC. Hellman and his colleagues will conduct the quantitative analysis and user interviews, develop the assessment tool, and participate in the writing of the white paper. Hellman is co-PI on the project but from an operational point of view he will participate in the project as a consultant.

Dr. Rupert Gatti is Fellow and Director of Studies in Economics at Trinity College, Cambridge, and a co-founder of Open Book Publishers which has published over 70 open access ebooks since its foundation in 2008. Gatti will identify sample books from the OBP catalog, give access to usage information, include links to the assessment instrument in OBP open access ebooks published in 2016, provide advice on the development of qualitative measures, and participate in the writing of the white paper and recommendations.

Rebecca Welzenbach is Director of Strategic Integration and Partnerships for Michigan Publishing, the publishing division of University of Michigan Library that includes the Press. She is a librarian with deep experience of open access book publishing. Welzenbach will identify sample books from the UMP catalog, give access to usage information, include links to the assessment instrument in UMP open access ebooks published in mid-2016, interview information supply chain intermediaries, and participate in the writing of the white paper.

Gary Hall is Professor of Media and Performing Arts at Coventry University and is a cultural and media theorist working on continental philosophy, cultural studies, new media technologies, and digital humanities. In 2006, working with Sigi Jottkandt, David Ottina, and Paul Ashton, he founded Open Humanities Press in response to the perceived crisis in academic publishing. Hall will inform the design of the qualitative study and will contribute to the white paper.

Ken Varnum is Senior Program Manager for Discovery, Delivery, and Learning Analytics at the University of Michigan Library. He will assist with the preparation of usage data from Michigan, provide input during the research, and participate in writing of the white paper.

Raymond Yee is an independent technologist. He is author of the leading book on web mashups, Pro Web 2.0 Mashups: Remixing Data and Web Services. He has taught classes on Open Data and Data Remixing at the UC Berkeley School of Information. Raymond served as the Integration Advisor for the Zotero Project and managed the Zotero Commons. Part of the founding staff at Unglue.it, he directs operations at the Free Ebook Foundation, and will assist with technology development.

Audrey Evans is research librarian with strong analytical, administrative, and project management skills. She received her MLIS from LIU in 2012. Most recently, she was Head of Research for Dollar a Day. She will conduct qualitative interviews.