Charleston Library Conference

Data Expeditions: Mining Data for Effective Decision-Making

Ann Michael
Delta Think

Ivy Anderson
California Digital Library

Gwen Evans
OhioLink

Ann Michael, Ivy Anderson, and Gwen Evans, "Data Expeditions: Mining Data for Effective Decision-Making" (2018). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/https://doi.org/10.5703/1288284317002

Data Expeditions: Mining Data for Effective Decision-Making

Ann Michael, Delta Think

Ivy Anderson, California Digital Library

Gwen Evans, OhioLink

The following is a transcription of a live presentation that was given at the 2018 Charleston Conference on Wednesday, November 7, 2018.

Ann Okerson: I've been asked to introduce the next session on "Data Expeditions: Mining Data for Effective Decision-Making," and I think it is actually a very good follow-on to what we learned from Annette Thomas. In this session we are going to hear from library experts about their scholarly publishing data hunting expeditions and the innovative ways in which they access and utilize deep data to inform their discussions and decisions and support their activities.

So, using data to inform decisions is a hot topic, but how is it actually done? And how can libraries and consortia find, manage, and leverage data in many of their activities, not just limited to publisher negotiations? So, in this session we are going to hear from several people who work hard in this area: Ivy Anderson from the California Digital Library talking about analysis that they do, not just for negotiating but for journal reviews and for publication analysis purposes as they transform from subscription to publishing support. Gwen Evans, the executive director of OhioLink, is going to present case studies on how OhioLink uses data to change their perspective on issues and find new ways to address them. The convener of this panel is Ann Michael, who is CEO of Delta Think, and she's going to introduce foundational tenets of using data to inform decisions, and she's going to moderate questions with the group, so I'm handing over to Ann, I would say "other Ann," but I think that's me.

Ann Michael: Thanks. Hi, and actually from a foundational tenet perspective, I really just want to put a couple of thoughts in your head to have you thinking about things as you are listening to Ivy and Gwen. And Ann mentioned one and the first thing is we talk about data to inform decisions, not data to make decisions or, you know, everyone talks about decisions that are driven by data, and I think it's really important that we understand that data is a tool and that it is a tool we need to use with skill and with sometimes a degree of finesse, so I hope you keep that in mind as you listen to the speakers. The other one, sorry, this is just a little too high, I'm too short. The other one is a concept that we hear all the time, which is not to let the perfect be the enemy of the good, and this is especially true in data. I've seen time and time again that in an effort to find the perfect data for something, this is something that dots every "i" and crosses every "t," we leave a whole lot of value on the table, value that comes from estimation and modeling the data we have, recognizing where it might be lacking and then compensating for that in other ways. So, again, thinking about that as a concept, that you work with the best that you have. There is a famous statistician, George Box, who says that "all models are wrong but some are useful," and I think that is a really good way to look at data in the sense that all models, no model is perfect. And then finally, one thing I think you're going to hear a lot of in what Ivy and Gwen have to talk about is data as an asset. Like any asset it requires time, it requires investment, it requires a special skill set, but also once it is made consumable as many other assets, it can be used by many people in an organization. So, with no further ado, we're going to start with Ivy and be thinking about your questions. We're going to leave time for questions at the end. Thanks.

Ivy Anderson: Thank you, Ann. So, just a little bit about the California Digital Library. I think many of you know CDL fairly well, but just to set the stage: the University of California is a 10-campus system; CDL is sometimes called the 11th university library. We were formed in 1997 to support digital library services for the entire University of California system, and we do work very much together as a system. We've had a mantra for many years (although it's gone in or out of favor at various times): "One University, One Library." And we try to make decisions collectively as a group in many of the areas that I'm going to talk about today.

So, I'm going to talk about two primary use cases for some of the work that we're doing with data. One is journal value analysis and decision support for journal cancellation and retention decisions; and then

https://doi.org/10.5703/1288284317002

the other set of use cases that I'll talk about is some of the newer work we're doing around open access modeling and transformation. The University of California system is very focused now on open access transformation, and we're doing a lot of data analysis to support our work in that area.

First, let me talk about our journal value analysis. A number of years ago, we turned our attention to how we might apply a more rigorous approach to our decision-making for journals and journal packages. Journal packages are a very significant part of our collective licensing activity at the University of California. We've often over the years deployed groups of librarians across the system to make decisions about which titles to retain collectively in a given journal package as well as which titles to cancel; the libraries spent years trying to determine whether there is a core set of journals that we should be licensing together as a system. But in reality, a lot of subjectivity is applied to those kinds of decisions. While we assembled numerous data points such as usage, cost per use, impact factors, and so forth, these were all treated as individual data points; we'd hand our librarians a big spreadsheet and say "Okay, have at it and then tell us what we should keep and what we should retain." When it came to journal packages as a whole, we generally had no way of relating the value of one package to the other; each negotiation was its own de novo activity.

So, a number of years ago we decided to take a more holistic approach to evaluating our journal packages and titles. To do that, we developed a more comprehensive metric that didn't just look at usage, which, while it is a very important measure—and COUNTER has done a lot of important work to try to normalize and standardize the way usage data is counted—there are still many factors that distort usage data and make it problematic as a sole measure of value; and we also wanted to find a better way to bring a variety of indicators together. So, we developed an algorithm that looks at journal value from three different perspectives:

1. We evaluate journals from the perspective of *Utility*, by which we mean both usage and also the citation behavior of our users—how often are our faculty and authors citing work in different journals?
2. We also look at *Quality* measures such as impact factor and Source Normalized Impact per Paper, or SNIP.
3. And finally, we look at *Cost-Effectiveness* measures—specifically, cost per use and cost per citation.

Then we roll up that data according to an algorithm that we've developed and apply it to every journal that we license across a range of 160 or so subject disciplines, to develop a broader picture of the value of each journal across our entire licensed portfolio.

This produces graphics like the one on this slide, which depicts the overall value of each journal package that we license based on the value of the specific journals within that package. Each journal is assigned a numerical score, and then we group these scores according to a range of values from high to low. This scoring helps us decide what journal packages we should target for value improvement or for potential cancellation, vs. which are already providing strong value. The pie charts that you see on this slide show that within any package there is a distribution of high-value vs. low-value titles; you can see that on the left-hand side of the chart, the packages that are of higher overall value have a large number of journals with very high value, whereas the lower value packages on the right contain many more journals in the lower value tiers. So this data allows us to evaluate our journal holdings both at the package level and at the individual journal level.

We can look at this data from a variety of graphical perspectives. So, here you're seeing a chart that shows the value of a given publisher's journals across a range of disciplines. The orange line depicts the value of that publisher's journals in each discipline compared to the average value of all of the journals we license in each of those disciplines, which is depicted by the gray line. What you see here is that for this particular publisher, while their journals in some disciplines provide better than average value to our community, most of their journals are below the norm in terms of value; and this gives us a basis for negotiation with a publisher in terms of how to improve the financial value of the package, and/or in terms of title-level cancellations.

Data at the journal level is very useful for things like title swapping when journals transfer in and out of packages. Having a numerical score for each journal allows us to rank journals within a given discipline, and this helps with decision-making. Decisions are still made by our librarians, so we are not using this data mechanistically, but we're using it as sort of a "first cut." As Ann said, it is data to inform decision-making, giving us a more holistic and objective picture of journal value across our packages.

We've also done some regression analysis recently to establish pricing targets for journal packages when these measures suggested that the package value was out of whack. So, we have done some analysis to correlate, for example, the pricing of our packages with a variety of quality measures; where we've been able to identify a correlation, we've used that to establish a pricing target. This chart gives you a depiction of those results: you can see that there is a band where most of our packages cluster in terms of a correlation between quality measures and pricing, with some packages that provide better value on the lower left and others that provide poorer value where the pricing is actually higher than our correlations say it ought to be; and we've been able to talk with publishers about how to bring the pricing more in line to produce a better correlation of value with pricing.

I also want to say thank you to those publishers in the room today who have worked with us in these areas; because we have worked with a number of you on ways to improve value, and in many cases these discussions have helped us to find ways together to retain journals in our licenses.

I'd like to move now to our OA transformation activities, which is an area that we're really excited about these days. I don't know how well you can read text on the slide, because I can't actually see it myself from here—but in these analyses, we're trying to look at our publishing output from a variety of perspectives. Something that is very important to the UC system right now is to better understand how we might move the subscription system toward open access; and in order to do that, you really have to understand the publishing behavior of your own community in a very deep way in order to understand the financial impact of supporting open access rather than supporting subscriptions. And so we've been doing a lot of analysis to understand the University of California's publishing output according to a variety of attributes. What is our corresponding authorship rate? How much of the research that's published at UC is grant funded? And what does this tell us about how we might be able to financially support the publishing output of our institution?

We're also looking at open access pricing measures. We've analyzed a lot of data in order to correlate our publishing output to publisher APCs; for example, to determine what articles are already being published open access to what we're spending for those articles on top of our subscription licenses. This is a hard problem for our libraries; we often don't know what our institutions are spending outside of our libraries for open access. To figure that out, we've mapped our publishing output to Unpaywall data, which can tell us which articles are published as Gold open access (in both Gold and hybrid journals), and we then mapped that back to publisher APCs in order to estimate what we're spending for open access on top of our licenses.

I won't spend a lot more time on this, except to say that it's one of the newer areas that we are working in and we're really doing some interesting work here. For example, this is a picture of how University of California publishing output is distributed across a range of publishers: 80% of our publication output is with just 25 publishers. This gives us a good sense of the publishers we need to target for open access transformation. We've pulled this data into a modeling tool—I'm assuming you can't read this very well, or if you can you probably can't tell what it means because it's very hard to explain in a single slide—but this is just a little teaser to give you an idea of how we're trying to bring a variety of data to bear to model open access transition scenarios. By compiling information about publication output, the cost of APCs for various journals, the number of articles resulting from grant-funded work, and so forth, we can model scenarios that allow us to input variables such as level of APC discount, level of grant-funded support, and so on, to model what a flipped world might look like from a financial perspective. We're continuing to improve on these modeling tools as we work with them, and we're very interested in trying to help the broader community to undertake this kind of analysis as well. We've been talking with a number of other institutions and libraries about how we might transfer the knowledge that we're developing in this area and make this a more generalizable tool for the community. This is an area that we are very interested in right now, and I would be happy to talk with folks about it later on.

Ann Michael: Thanks, Ivy. Now let's move on to Gwen and then we will have some time for questions.

Gwen Evans: Good morning. I am Gwen Evans, executive director of OhioLink, and I'm here to talk about viewing data differently, and I'll present two very different data problems, one highly static and one very dynamic, and how the lens through which you view them makes all the difference.

Our organization, again like Ivy's, is a membership organization. We are a state agency of 90 institutions, and that includes all public higher education institutions in Ohio as well as almost all independent colleges and universities of any size, the State Library of Ohio, and special focus institutions both large and small. Having such a diverse range of institutions creates a variety of data challenges in itself as the meaning and meaningfulness of data varies widely across the membership, and I just want to emphasize again we don't at OhioLink make the decisions. We help our membership make the collective decisions that work for them.

So, this is just a brief, a partial, it's not brief, it's a partial list of our services to give you some idea of the scale, and the two that relate to my examples are we run a print sharing and delivery network with a Central Union Catalog, and we help administer five regional high-density storage facilities for low-use print. We negotiate and contract for approximately $30 million annually in shared digital content and do a lot of collection budget and financial analysis across our 118 libraries. In our latest initiative, addressing textbook affordability, we negotiated statewide pricing agreements with the major commercial textbook publishers, so a huge part of our core competency and value as an organization therefore deals with managing data at scale, metadata ingest, data normalization, usage analysis, financial and budget information, as well as the very important issue of presentation and visualization of complex data in order to explain what is happening both to subject matter experts and to lay audiences.

So, my first example is drawn from a very traditional library endeavor, managing print holdings in limited space. So, we coordinate five regional depositories, which collectively hold over 8.5 million low-use circulating monographic and serial volumes. Each depository stores the materials in a Harvard-style high-density storage facility, 40-foot-tall stacks, shelved by size, retrieved manually using a modified forklift. They are full to all intents and purposes. These were never designed to be last copy repositories. They were designed and filled as extensions of each individual depositing library, just cooperatively managed and shared, thus there was no attempt to de-duplicate on ingest of the materials. A 2013 OCLC research study estimated that of the system-wide print holdings of our depositories, 75% of the titles are duplicated in more than 99 World Cat Libraries.

So, the budget for operating and maintaining these facilities comes directly from the Ohio Department of Higher Education. With the aging of the facilities, the reduction in demand for print, and the high rate of duplication, the cost per retrieval is increasing at an alarming rate, and make no mistake, in a state agency environment these sorts of data and cost calculations are absolutely asked for when we justify our budget requests. So, our dilemma is how to make the space more valuable instead of presiding over ever decreasing use at an ever higher cost under the scrutiny of frugally minded state administrators.

So, the de-duplication of Harvard style—it's not really cost-effective to de-duplicate, as many experts have pointed out, so the obvious solution is to recoup space to be refilled in a more effective manner. However, removing an item because it is a duplicate leaves a gap in a fixed order and in order to recoup this space every weeded item automatically invokes touching many other items both physically and within the database. So, what to do? Instead of focusing on de-duplication, which we've been talking about for years, OhioLink and Ohio University, which runs one of the depositories, decided to redefine the problem using the same data. So, we use a change of perspective. What is the minimum set we actually have to touch to recoup space? But the focus is de-duplication in the depository monographs. That's more than 4 million items. If the focus is uniqueness, that's a mere 500,000 items and that changes your risk profile, because of changes over time, Mark standards, local cataloging policies, lack of controlled vocabulary, especially at the volume level, there is uncertainty in the dataset. This messiness in catalog data leads to a false positive identification as "unique" if you're focusing on uniqueness but that errs on the side of preservation of the scholarly record to our benefit. At this scale it's unnecessary to worry about duplicates that are inadvertently kept as unique. A focus on de-duplication, on the other hand, requires management of the risk of inadvertently discarding unique items, which is much more of a dire risk with much more work and double-checking involved for many more items.

So, OhioLink and Ohio University, which manages the Southeast Regional Depository, wrote a joint grant proposal to test the idea of compressing an entire depository by focusing on uniqueness instead of duplicates. OhioLink staff unsuccessfully tested the concept of using the OhioLink Union Catalog to identify unique monographs, but we don't have the resources in terms of people or software to compare it and identify titles at this level. However, OCLC's sustainable collection services, on the other hand, did have the capacity to provide data analysis at that level. So, OU, with the help of the GreenGlass tool from SDS, eventually identified 60,000 unique items, monographic items in the depository. These were moved out either to the main library or to at least storage facilities. The next step will be to identify and remove unique serials, which will be a different data challenge. Are you asking yourself, "But you still have to touch all those duplicates?" But how we touch them has implications for cost and time. At a certain point what remains in the building after unique items are removed will be discarded en masse. OU's experiment and focusing on uniqueness instead of duplication may show us a potential way forward for some of our depositories to either recoup depository space in a single depository or by sunsetting some of the depositories by relocating unique items into the others. So, our newest data problem revolves around OhioLink's negotiations at a statewide scale with commercial textbook publishers. This is very unfamiliar ground for us and our biggest challenges were data challenges, both at the beginning of the process and now.

Initially, we took the standard consortial negotiating approach, which is find out what everyone is already using and define that as the target collection. The last thing you want with textbooks is to acquire content that might be assigned. That's simply not how faculty work. You really need to know what is already in use. Our parent agency collects massive amounts of data from the public institutions already, and it seemed as though we could just ask them to send that with the other data. But it turned out that institutions can't even get their own data about textbook assignments in an easily analyzed aggregate form. Faculty create the desired metadata about textbooks in proprietary interfaces owned by bookstores in highly decentralized processes. Often bookstores don't have to provide that data to their own institutions because it is considered competitive business information. That data is almost never exposed in an easily collected online format for similar reasons. So, even if some of our institutions could and would supply that information, trying to get it for more than 30 separate campuses from the publics would pose severe challenges in timeliness, data normalization, and even figuring out who to ask for the data. What students actually would pay, of course, is a completely different data problem and incredibly dynamic. We have no way to know what they're paying when they are independent buyers on the open market. What are they paying at their campus bookstore for new print? What are they paying used from Chegg? What are they paying as a digital rental from Amazon? So, we eventually just define the collection as everything from the major textbook publishers: Wiley, McGraw-Hill Education, Macmillan, Sage, Pearson, and Cengage. Our inability to get and manage the kind of data we were used to crunching and learning instead to live with the data we could get influenced the model and agreements we eventually settled on.

Our second biggest challenge was and is making sure that the prices we negotiate are advantageous to students, but, given that retail markup at campus bookstores is variable and there's a national online market that students use, we needed data analysis tools that just weren't in the library toolkit. So, in order to monitor our pricing, we are simulating being a college bookstore in order to gather business intel. We're using a product called Verba Connect, which bookstores actually use to make sure that their pricing is competitive on the market. We're not selling content, but we do have an OhioLink price that we want to ensure is competitive with other readily available sources of textbooks, and I want to emphasize we've been using this for about three weeks, so we're still figuring out what the limits and capabilities are, and we are using it in a way that it was not designed to be used, so analyzing aggregate publisher data, which we are interested in but bookstores are not interested in, will require some hacks that we will have to do in-house, but you can see here and there that we can check our prices against the major online textbook market prices in a variety of formats.

This tree map of our prices versus market prices exposed one of our most dynamic aspects of this data. Cooler, bluer tones, which cluster in the lower right, indicate that the OhioLink price is beating the online market. Warmer tones, which cluster in the top left, indicate we did not do a very good job of negotiating. When I first saw this tree map, I was like "Get me the phone! I have to have some words with the publishers." However, Verba pointed out that right now we're in the middle of the academic semester, and this is the price online in national markets, probably the lowest it will be during the academic cycle. About two weeks before the semester begins demand starts spiking and so will prices, and if you don't think Amazon in particular doesn't indulge in surge pricing, think again. So, now we have a data scheduling protocol to follow in order to analyze if our deals are actually competitive or not.

So, I have some last thoughts on data. The ability to collect, organize, analyze, and manipulate data has always been a fundamental competency that libraries fostered. My first example, unique versus duplicate, relies on very familiar kinds of library data, but recasting the target population for action within that data set resulted in a potential solution to what had seemed to be an intractable problem for us. Increasingly, however, the data we need is no longer solely under our stewardship, as Ivy pointed out. We are operating at vaster scales or our data has to be combined from disparate sources. In my own organizations our hiring has increasingly reflected a need for a sophisticated and systematic approach to data analysis, whether it's hiring someone whose main job is data analysis and management across the organization or defining data analysis and management as a core competency in more and more job descriptions. We are using more consultants or purchased data analysis, services, and tools. We consider the data, the quality of it, as well as can we manage and analyze it as an integral part of the assessment of any new service. I remarked yesterday that as a consortium we only exist in the aggregate, so data is a fundamental way that our existence as a consortium is expressed. And that's—thank you for your attendance and attention and I'll happily answer any questions.