Cornell University Library
E-usage Analysis Task Force

Final Report
January 31, 2013

Task Force Members:

Adam Chandler (chair)
Rich Entlich
Zsuzsa Koltay
Mary Beth Martini-Lyons
Liisa Mobley
Table of Contents

Executive Summary

1. Introduction
   • Charge for the E-usage Analysis Task Force
   • Trends in libraries

2. Demographics of use and patron privacy
   • Demographics
     o Licensed electronic resources
     o Non-licensed, CUL hosted digital content
     o Print circulation
       ▪ Voyager
       ▪ Borrow Direct
       ▪ Interlibrary loan
   • Patron confidentiality

3. Inventory
   • Data sources and analysis framework

4. Data classes and use cases
   • Cost per use: what is possible and how much effort is required?
     o Licensed ebooks
     o Licensed ejournals
     o Licensed databases
     o Print
   • Patron satisfaction
   • How users find and consume CUL content

5. Conclusion

6. Bibliography

7. Appendices
   • Data sources inventory
   • Table: Overview of data sources
Executive Summary

Our report begins with a review of the literature on web analytics and data mining in libraries. The purpose of this review is to provide some background and context into which Cornell University Library’s current practices and policies may be situated. With this background we are able to construct an imaginary usage data management spectrum onto which all libraries may be placed. On one side of this continuum are libraries that do not actively collect or draw reports from what Dempsey (2006) calls the “intentional data” left behind by their users. On the other side of this line are libraries that treat data mining as a strategic goal. With the data compiled here, the reader will be better equipped to understand where CUL fits into this landscape.

Section two of the report dives into the details of current CUL usage data management practices, viewed through a demographic based lens. Our working assumption is the new Cornell University budget model will be a catalyst for increased interest about who is using our collections and services. We further assume that the unit of this analysis will likely be the 11 colleges that comprise the new budget model, though we might need to be prepared to provide reporting at the department level. There of course exists a fascinating tension between more detailed reporting on our users and the traditional library ethic to protect the privacy of our users. Links to our current privacy policies are included and an argument is made that it is time to review our print-based assumptions given our increasingly electronic collections and services environment.

The narrative in section three is short, and serves mainly as a pointer to the two appendices. One of these appendices is comprised of the full text of the data we collected on the different usage data sources we gathered into our inventory. These data sources range from obvious ones like COUNTER electronic usage reports and Voyager circulation to less well known sources, such as the Count It reference transaction database or an email troubleshooting list archive. Our other appendix is a spreadsheet that provides the reader with a higher level view of this data inventory.

Section four attempts to answer these common questions:

- Cost per use: what is possible and how much effort is required, for licensed ebooks, ejournals, databases, or print items?
- To what extent can we measure patron satisfaction and proactively identify frustrating service points?
- How do users find and consume our resources?

We hope the information presented here will educate the wider CUL community about the limits of and effort required for some ejournal cost per use requests, for example, while at the same time broadening the discussion to look at other interesting and actionable questions.
Our conclusion is brief. Given the charge, we felt it appropriate to use our energy to compile the most comprehensive inventory possible in the time we were given for this project. We assume and hope that library management informed by the data presented in this report will be motivated to take further action where they see it is appropriate. With that said, there are two themes, or perhaps one theme with two parts, that we feel need to highlight. The usage data that are being collected across the CUL landscape is quite literally out of control. Cornell Library would benefit from placing all of our usage data collection streams under explicit managerial oversight. Second, and this item is closely linked to the first, we need to review our privacy and confidentiality policies and align them closer to the reality of how usage data that is by way of our services being collected and stored about our users in 2013.
1. Introduction

Charge for the E-usage Analysis Task Force

This group is called together for a 4 month project to survey all existing data sources and questions they could answer by themselves or in some combination about the use of CUL’s digital services and electronic and print collections. The nature of the work is a broad survey, rather than detailed analysis of any one specific area. The priority of the group is creating an inventory of electronic resource usage streams (80% of effort), but print usage data should also be summarized (20% of effort) so as to provide a comprehensive picture across CUL. The goal is to better equip CUL to assess our investment in e-resources and e-services, such as:

- What does our data tell about disciplinary use, strengths and weaknesses? (The ability to add demographic understanding of the use of our resources)
- How do we tell we are making the right types of investments in collections? (The ability to have cost per use data on the title level for, e.g., collection development or cancellation projects)
- How do we tell we are making the right types of investment in digital services in support of electronic and other resources?

The team will work with relevant subject experts and staff as needed to:

1. Inventories all sources of system generated data about how and how much users connect to and use our electronic resources and services, drawing a clear distinction between what is already available to CUL and what could be available with additional investments in the future (possible examples include monitoring click-throughs, capturing and analyzing network flow information at the Cornell border, uniform and comprehensive web site and web search usage analysis, and a full understanding of the information that licensed service providers can or could give us.);
2. Investigate what it would take to obtain the various types of use data;
3. Think through what kinds of questions the data would let us answer;
4. Identify a set of potential projects that require digging deeper into these data, including qualitative projects that would supplement the quantitative data, project a rough sense of resource needs for each project (high/medium/low resource need); and

Consequently the nature of this phase of work is a broad landscape survey, rather than detailed analysis of any one specific data/service area.

Xin Li, Dean Krafft, and John Saylor will serve as sponsors and resource to the team to support their work.
Trends in libraries

Interest in using usage data to improve services and justify library expenditures is increasing. Much has been written about using COUNTER reports to study the use of licensed electronic resources. Less well documented are efforts in libraries to utilize web analytics tools and other data management strategies to study user behavior.

The Google Analytics (GA) product receives the most attention in the library literature, starting in 2007 (Fang). Plaza (2009) wrote about using Google Analytics to study the frequency of return visits to a site. That same year the Library Corporation announced that they had integrated Google Analytics into their LS2 PAC library OPAC product (The Library Corporation, 2009). McMullen (2010) recommended the use of GA to public services staff. Turner (2010), echoing statements made from Fang, argued that crude aggregate numbers such as total hits per year are inadequate. He recommended GA to study a library website’s key performance indicators, which he defined as "visits, bounce rate (macro), conversion rate (macro), average pages per visit, average time on site, percentage of new visits, and other reports." The July 2011 issue of Library Technology Reports, dedicated to Web analytics in libraries, also recommended using GA and included a chapter on how to integrate it into library websites. Hesse (2012) described the use of GA to monitor traffic on the Illinois Harvest Portal. Ma (2013) recommended it to the readers of Technical Services Quarterly.

What becomes apparent from reading these articles about Google Analytics is that a very wide spectrum exists across libraries. On one side of this spectrum are libraries that are doing no analysis of their website usage (the target for audience for these articles about Google Analytics, presumably). In the middle are libraries that are already using or starting to use Google Analytics or a similar web analytics tool. On the other end of this spectrum are a tiny number of research libraries that are assembling comprehensive data warehouses which can gather much more than is possible with GA.

One of the earliest discussions of the potential of a data warehouse for libraries is Su and Needamangala’s (2010) article about their Library Decision Support System method. Even though the model they describe is really narrowly focused on mining the usage data generated by a library management system, it was a start. In contrast, Nicholson’s “bibliomining” framework (2003 and 2006) is more comprehensive and far reaching in its design. Like Su and Needamangala, Nicholson brings data warehouse storage and query techniques into the library context, but Nicholson goes a step further. His framework explicitly uses methods that honor the traditional library ethic of patron privacy. In Nicholson’s model, transaction data that can be linked to a patron, whether
the key is the patron id in a circulation record or the IP address of a desktop computer, are removed and replaced by a group identifier before the data is stored in the data warehouse. It is worth noting that patron privacy does not appear to be a concern held by the authors cited earlier who advocate for the use of Google Analytics in libraries. Google Analytics routes everything to Google’s own servers, including IP addresses. The Google Analytic advocates also ignore the fact that this tool is not capable of linking usage to campus group demographics. Google Analytics links to individuals. Given the sophisticated tracking technologies used by Google, it is reasonable to imagine that the data collected by way of library sites is associated with data collected when the same user is logged into Gmail.

Nicholson spent several years unsuccessfully trying to promote the use of his group focused, privacy sensitive framework in libraries. In about 2007 he appears to have given up on the effort, switching to “gaming” (as in computer or board games) in libraries as his primary research interest. There does, however, appear to be at least one research library building a system consistent with Nicholson’s biblimining concepts, the University of Pennsylvania.

The Penn Data Farm was started by Joe Zucca and colleagues about 12 years ago, according to Zucca (2003). The Penn Data Farm has expanded from trying to consistently capture electronic resource log data to data for print usage, in library use, Borrow Direct, etc. Now the Penn Data Farm architects are abstracting what they have learned into an IMLS supported event-based model called METRIDOC (Zucca, 2011). In a very real sense the Penn Data Farm is the example that represents the other side of the continuum.

Penn is not alone. Other libraries are experimenting with demographic-based data warehouses. The University of Wollongong Library (Australia), in collaboration with their campus Performance Indicators Unit, built a data warehouse they call the Library Cube, to join library usage data with students’ demographic and academic performance data (Cox and Janti, 2012). Britton and Renaud (2012) from the University of Miami described a project with their Center for Computational Science, in which they link usage to departments. They also explored the correlation between library use and student GPA (the correlation in their study is weak). Staff in the University of Minnesota Libraries also explored the relationship between library use and GPA (Soria et al., 2013).

Another way of using demographic-based library data is to create ratios that compare user populations with some other variable. Lown (2009) created a database that matched fund code expenditures with the number of faculty and students in departments. His method reveals interesting patterns about how library collections are being targeted to constituent groups. Another example is Rich Entlich’s (2010) study that compared borrowing patterns by college.

Where is Cornell University Library on this data management continuum? We use usage data on a regular basis to make decisions, but the decentralized way in which it
is collected is far behind the leaders in this area. We hope this inventory will provide the reader with a clearer picture of where we stand on this scale.

2. Demographics of use and patron privacy

Demographics

The new Cornell budget model puts the emphasis on the colleges as the unit of analysis and accountability (Fuchs, 2012). As a result, there is anticipation within CUL that the Library budget may come under increased scrutiny by Deans. The January 14, 2013 CUL Managers’ Council minutes read, "Annual reports will likely go out to Deans to let them know what benefits their schools/colleges are getting from the Library. The Assessment & Communication unit can help unit libraries create these reports." Our task force charge stresses the importance of understanding "disciplinary use" of our library resources. How should disciplines be defined? The question at the heart of demographics of use is, at what level (individual, department, or college) does LibEx want to justify the CUL budget? According to Lee Cartmill, the following 11 academic units will be used in the new CU budget model (note, Weill and the Tech Campus are obviously also of interest to us here in the library, but they will continue to have separate budgets).

College of Agriculture and Life Sciences
College of Architecture, Art, and Planning
College of Arts and Sciences
College of Engineering
School of Hotel Administration
College of Human Ecology
School of Industrial and Labor Relations (ILR)
The Faculty of Computing and Information Science
Cornell Law School
Samuel Curtis Johnson Graduate School of Management
College of Veterinary Medicine

Given the obvious privacy violations associated with tracking at the individual level our task force did not explore that option. However, as discussed by Collins (2012, p.103), the Assistant Project Manager for Borrow Direct at the University of Pennsylvania, though we have no need, nor desire to maintain archival records of library users’ transactions at an individually identifiable level, there is considerable value to being able to track unique user actions. For example, it is one thing to know that x number of article downloads from a particular platform during a certain year were made by faculty members, or even faculty members affiliated with a specific college. It is another to be
able to say that \textit{x} number of \textit{unique} individuals of a certain status used that service. The latter form of data allows us to determine whether detected use is all from one very enthusiastic person (with others of the same status making no use at all), or if the use is more evenly divided over a much larger proportion of the total user population. In other words, do we have a narrow user base, or a wide one? To make such a distinction, it's necessary to initially track use with unique identifiers, but then to anonymize them so they can't be traced back to the individual. Decisions can then be made about what level of demographic information to associate with the anonymized IDs.

This brings us to a second issue, which is determining an appropriate level of granularity to use in demographic tracking. If the only issue were justifying library expenses to the deans, the eleven academic units in the Provost's budget model might suffice. But there are other considerations. Selectors typically deal with department heads or departmental liaisons. When resource costs such as for multidisciplinary databases or ejournal packages are divided among selector funds, it is department level use and benefit (or the perception of it) that forms the basis for the decision. For that purpose, usage data at the department level would be more valuable than that at the college level.

Furthermore, use data at the department level can lead to better decisions by selectors about what resources to purchase, and how to fine tune the contents of packages. It can reveal whether the expected users of a particular resource are the actual users, and provide evidence for cross- and inter-disciplinary interests they might not otherwise be aware of. Selectors could also learn which resources are preferred by members of which departments, and then tailor the contents of those packages accordingly. This makes for both better resource utilization, and more responsive resource selection.

There are two major concerns about gathering and maintaining department level use data. One is how much effort is required, and whether the necessary data and infrastructure even exists to accomplish it. A lot depends on what kind of data about Cornell's internet traffic and the assignment of IP subnets we're able to get from CIT. If they're willing to provide the data, tracking use at the department level would not be technically difficult.

The other concern stems from our promise to users not to monitor their "reading habits" or to compile records of them. Demographic units that are too granular can unintentionally unmask individual users (O'Rourke, 2012). Thoughtful planning and execution of user data management can mitigate such breaches. We have the option of aggregating overly small departments, or we can use less specific indicators of the materials being used. For example, if we had data that tracked downloads of articles at the title level, we might want to associate it with demographic units no smaller than the college, or large aggregates of departments. On the other hand, for use data at the level of the platform (for example, ScienceDirect, or Sage journals), it might be acceptable to associate it with demographic units that included small departments, because the content being used is described in only the vaguest terms. In that case, even if there were, hypothetically, a department with only a single member, would we
consider it a breach of privacy to know that the person was a user of ScienceDirect? (Since April 2012, we've had the ability to associate circulation of individual titles in Voyager with the user's field of study, for students in the graduate school). What level of content description is appropriate to associate with what level of user description? This would be a good issue to revisit as we think about revising our confidentiality guidelines for a more e-resource intensive era.

To what degree are we prepared to map the use of our resources to the 11 college units (or to departments, patron status, or other demographic measures) in the new budget model? The answer depends on a number of factors, mostly having to do with the format of resources and where they are hosted.

**Licensed electronic resources**

When we sign a license for an electronic resource, the license usually includes the following definition for the Cornell University community:

> Cornell University consists of 4 campuses and numerous offices throughout New York State and beyond. All of these report to the same administration on our main, Ithaca, NY campus. These campuses all have dedicated IP address ranges.¹

The function of this definition and essential task of the management of licensed electronic resources is to draw a distinction between the Cornell community and the outside world. Given the current technology available across libraries and vendors, the mechanism used to enforce the subscriber/non-subscriber distinction is the IP address of the user's device. Each time we license a new resource we send CU's official IP range list to the vendor. Only requests coming from these IP addresses are allowed in by the vendor. In terms of usage data, everything else that follows is optimized around the binary subscriber/non-subscriber distinction. The IP address of the device is recorded in the vendor’s server log when a request is made and from there reports, such as those adhering to the COUNTER Code of Practice, are created and made available to the library. Occasionally vendors offer detailed reports by IP address range, but this is not usually the case, and these reports when they are available are not formatted consistently. What we consistently receive from vendors are aggregate numbers (sessions, searches, or downloads) for the whole Cornell community. Therefore, we do not at this time have comprehensive IP address level usage data from vendors. If a Dean in one of the schools walks into Lee Cartmill’s office tomorrow asking how much his students and faculty used the Library’s licensed electronic resources last year, we cannot provide an answer because we do not know.

With development effort, we could, however, provide a comprehensive session level report of requests made by way of library discovery systems to licensed content by a

¹ We also prefer to include the following statement in addition to the existing definition of Authorized Users: 'In addition, patrons not affiliated with Licensee who are physically present at Licensee's site(s) ('walk-ins') may access the Licensed Materials from within the premises. Walk-in users are not permitted remote access to the Licensed Materials.'
college or department. This is possible because the proxy server requires NetID authentication when the initial request is made. We prefix our proxy server to the host URL before loading the record into Voyager, A-Z lists, etc. The purpose of this modification is so the proxy server gets invoked automatically if the user is off campus. So it would be possible to log requests made by a given NetID and through a batch process map the NetID to a college or department affiliation, then store the college or department name in a data warehouse along with the URL of the requested resource. The NetID at that point would be deleted to protect the patron’s privacy.

That solution covers access made through library discovery systems only, so it is not comprehensive, which might be a political problem. What about on campus activity made outside of library discovery systems, such as when a student on campus types in “jstor.org” and goes to it directly? Dean Krafft is exploring the feasibility of collaborating with CIT on a project focused on the data CIT captures in their perimeter routers. This approach would, in theory, provide us with department or college level IP usage for machines with static IP addresses such as desktops. These reports would not be as precise as the NetID based reports described above. There is also wireless activity to consider -- it is unclear to the authors what kind of reports are possible from on campus wireless devices using this scheme. Given that owners of wireless devices are required to register them with CIT, presumably CIT could mine that data in the same way we could mine our proxy server data.

We are, however, not completely in the dark about who across the CU community is using licensed electronic resources. Though not comprehensive, we have valuable sources of data from which to draw. For requests made through Cornell discovery systems, a report available in the CUL Logs system provides location based reports about the use of our licensed electronic resources. As an illustration, these are the top 10 locations from which licensed electronic resource requests were made on January 29, 2013:

<table>
<thead>
<tr>
<th>City, State (LocType)</th>
<th>Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wireless (CU (Campus))</td>
<td>1609</td>
</tr>
<tr>
<td>Ithaca, NY (Ithaca not CU)</td>
<td>1123</td>
</tr>
<tr>
<td>Veterinary Medicine (CU (Campus))</td>
<td>197</td>
</tr>
<tr>
<td>El Paso, TX (USA not NY)</td>
<td>166</td>
</tr>
<tr>
<td>China (Overseas)</td>
<td>152</td>
</tr>
<tr>
<td>Arts &amp; Sciences (CU (Campus))</td>
<td>131</td>
</tr>
<tr>
<td>LIB: Mann: staff (CU Lib (Staff))</td>
<td>117</td>
</tr>
<tr>
<td>LIB: Olin: staff (CU Lib (Staff))</td>
<td>115</td>
</tr>
<tr>
<td>Weill Medical College in NYC (CU (Weill))</td>
<td>112</td>
</tr>
<tr>
<td>Engineering (CU (Campus))</td>
<td>104</td>
</tr>
</tbody>
</table>
Non-licensed, Cornell hosted digital content

Each year CUL dedicates significant resources to building and maintaining digital collections and library web sites. To what extent do Cornell patrons make us of them? Sites hosted in the CUL Logs system (http://logs.library.cornell.edu) have detailed demographic reports available today, with data reaching back five years in some cases. Our local system was developed over the course of 2007-2011 (Chandler, 2012). It is based on the bibliomining theory developed by Nicholson. A detailed explanation of the Cornell system components is available at: https://confluence.cornell.edu/pages/viewpage.action?pageId=109228183

For the purposes of this report, the important part to understand is how the location function works. When a log file is processed in the system, the Apache log file row is parsed. Here is an example of a log row:


The first part of the line in this example, 127.0.0.1, is the IP address. The Logs system takes this address and attempts to match it against one of three IP lookup tables. The first lookup table it consults contains the IP address ranges for CUL desktops, both staff and public. This table is maintained by Linda Miller in the Assessment and Communications department, with assistance from Desktop Services. If no match is made in this lookup table a second table is consulted. The source of the data in the second table is CIT: they provide IP address range information for 23 academic units, 18 administrative units, and 2 affiliated units across the CU campus. The third IP table is a data set purchased from a company called Max Mind that gives us geo-located data around the world. With all of this in place it is easy to learn, for example, that about 4% of the users of collections hosted on our DLXS digital collections server (http://digital.library.cornell.edu) server are affiliated with Cornell (most of whom are using wireless devices).

Recently we discovered that Library Systems has shifted development resources away from the CUL Logs system to an application called AWStats. Like Google Analytics, AWStats does not offer campus-based location usage reports. Also, similar to Google Analytics, AWStats retains individual IP addresses.

Print circulation

Voyager: PeopleSoft/Workday demographic information:

Although we think of the Voyager circulation module primarily in terms of print resources, we in fact use it to track use of most of the physical resources in the library's
circulating collections, including other information formats such as CDs, DVDs, and microforms, plus convenience items, such as laptop computers, AC adapters, digital cameras, and even study rooms. While in use, items are charged to individually identifiable patrons (or to carrels, studies, or other institutions, that can ultimately be traced to an individual), so that we can make sure the item is returned, or if not returned, paid for by the borrower. In order to protect patron privacy, our Voyager installation is configured to sever the connection between the individual borrower and the item borrowed at the moment it is discharged. The only information about the borrower that we permanently associate in the Voyager circulation archive with the borrowed item is their patron group (e.g., faculty, undergrad, grad, staff, etc.)

However, Voyager has another means, outside of the circulation archive, to track the demographics of materials use. Until recently, we've made very little use of this facility (based on the PATRON_STATS table), but in April 2012, we loaded a set of codes for graduate fields of study into Voyager. As a result, Voyager now maintains a permanent record that associates a particular circulation transaction with the graduate field of the borrower. Although the association is maintained in a different table, the effect is the same as adding the graduate field of study to the circulation archive, alongside the patron group information that we've long retained.

The data to make the association between patron ID and graduate field comes from the PeopleSoft/Workday patron feed. That file also contains information about the college affiliation of undergraduates, the college affiliation of faculty, as well as the departmental affiliation of faculty and staff. So tracking use of items that circulate through Voyager is possible for all of those demographic categories. Rich Entlich has requested that the pseudo-patron IDs associated with our Borrow Direct partners be input into PATRON_STATS. See the Borrow Direct section for more details about the intent of that request.

For the past three years, Rich Entlich has been collecting demographics about the users of print monographs and music scores in Voyager through the use of circulation snapshots. The snapshot is a record of all items in circulation at the moment the data is extracted. Since the items are currently in use, Voyager knows who the borrower is. Once the snapshot is taken, it is enhanced with demographic data from the patron feed. Then the patron IDs are carefully anonymized, but their unique character is maintained. This allows analysis about not only which user types are borrowing which material types, but how many distinct individuals are doing so.

Voyager also has the capability to maintain a parallel set of data about the items in the collection. Particular items could be associated with specific events that might have an impact on use, such as moves between unit libraries, or to the annex. The system would record not only the physical movement of specific volumes from one holding location to another, but the timing of those events. If we were recording such data, we'd have the ability to determine whether there are correlations between an item's circulation and where it is housed.
**Borrow Direct:**

Because BorrowDirect (BD) transactions involve two institutions, the transaction data for them winds up being somewhat fragmented. The loaning institution’s ILS will have complete bibliographic data about the transaction, but may know nothing about the borrower other than, perhaps, that he or she obtained the item through BD, rather than traditional ILL. Our Voyager installation didn’t start making that distinction until August 2010. Prior to that point, all we knew was that a transaction involved a borrower via some form of ILL.

As mentioned earlier, since we used fixed pseudo-patron IDs for each of our BD partners, we could add the name of the borrowing institution for each BD loan transaction if we added those IDs to PATRON_STATS. This has been requested, but not yet implemented.

The other source of demographic data for BD transactions is the Penn Data Farm. The BD transaction data in the Data Farm is gathered in real time by the Relais system, and includes the name of borrowing and lending institution, as well as a patron status. Unfortunately, the system does a poor job of capturing certain bibliographic data, and is missing some critical elements such as language of publication and OCLC number. Penn staff have told us that the system is too fragile to support extraction of additional bibliographic details.

An effort is underway in BD to add a unique identifier, such as an item ID to each transaction in the data farm, which would allow the missing bibliographic data to be filled in from our own ILS. If successful, it would allow us to associate specific items in our collection with both the borrowing BD institution, and the status of the borrower at their home institution.

However, other than through circulation snapshots, which could miss a lot of BD transactions, given that the loan period is only for six weeks, and we take snapshots only eight times per year, we have no means of determining the number of unique individuals who are using BD.

**Interlibrary loan:**

The Illiad ILL software we use includes complete data, including NetIDs, for all transactions made since 2000. To our knowledge, there is no analysis being done on this dataset, but much could be done: it is a rich resource. NetIDs are retained in this transaction history. CUL should make a decision about whether this dataset has value. If it does have value, the data should be exported, mapped to appropriate demographic markers, then purged. If it does not have perceived value it should simply be purged.
Patron confidentiality

Protecting patron privacy and confidentiality has been an important part of the professional ethos of librarianship for a long time. Moreover, New York State law protects the confidentiality of library records related to the use of materials and services. See CUL’s public statements of these principles (Cornell University Library, “Confidentiality”; Cornell University Library, “Privacy and Confidentiality in the Cornell University Library”). In the aftermath of 9/11 it became a widely accepted best practice to minimize the retention of data and purge library records of personally identifiable use information as broadly and as regularly as possible. CUL followed a similar direction (Koennecke, et al., 2003).

In the era of security cameras, Facebook oversharing, and data mining as an everyday corporate practice, fewer and fewer people seem to mind the lack of privacy in everyday life, especially if there is something to be gained from the exposure, such as personalized services (Eler, 2012). “The privacy that libraries traditionally have been preserving is not always valued by their patrons, especially in an age of social networking,” says David Weinberger, co-director of the Harvard Library Innovation Lab (Parry, 2012).

At the same time, more than ever before, libraries rely on third party commercial entities for the electronic delivery of information. Their privacy practices cover a broad range from reasonable to the complete monitoring of user activities. Even CUL’s arXiv and Project Euclid retain complete log files permanently. CUL publicly identifies its practices but emphasizes the users’ role in familiarizing themselves with the practices of vendors (Cornell University Library, “Library Practices on the Collection, Use, Disclosure, Maintenance and Protection of Personally-identifiable Information”).

By 2013 privacy in libraries has become a balancing act, rather than an absolute principle (Parry, 2012). Retaining personally identifiable data without a business purpose that overrides privacy concerns (e.g., billing for titles not returned) should still be avoided. However, the wholesale dumping of use data that could be anonymized or aggregated for information that does not hurt privacy but helps the library with new services, decision-making or the telling of its story, should be reconsidered. The CUL Logs system provides a nice example of this approach, something that could be extended to other library-maintained systems, because it is explicitly designed to maximize the utility of the data being collected in transaction logs using methods that are consistent with your traditional ethos to protect patron privacy.

Partnering with other keepers of personally identifiable information on campus (e.g. CIT) could be done in a way that prevents looking over the users’ shoulders as they search for information, while it could still provide us with valuable data on the demographics of e-resource use, for example. Such decisions need to be based on case-by-case risk/benefit analyses.
3. Inventory

Data sources and analysis framework

The first step in compiling our inventory was identifying the various pockets of usage data that are being collected across CUL. Given the experience with these systems within our task force, this went quickly. For our study we identified more than a dozen different sources of usage data of varying types. A few of these data sources are aggregators which collect usage data from multiple sites; AWStats (120 sites), Google Analytics (18 sites), and the CUL Logs system (more than 50 sites) fall into this category. Other data sources apply to only one particular service: Borrow Direct, Count It, ezproxy, Illiad ILL, LibGuides, MyiLibrary, Summon, WorldCat Local (Omniture), Voyager circulation. COUNTER is the "code of practice" used by licensed content providers. "Vendor custom reports" is a cousin of the COUNTER vendor reports: these are reports offered by licensed content providers that offer information that is outside the COUNTER regulated report types. We also include public facing email troubleshooting list archives in our analysis. This is a little different than the other sources listed, but they do tell something important about our patrons’ experience with our collections and services. To make sense of such a wide array of data, we created a framework that includes four criteria.

<table>
<thead>
<tr>
<th>criterion</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic counts</td>
<td>session counts, page views, downloads</td>
</tr>
<tr>
<td>behavioral context</td>
<td>location/demographic, staff/patron, browser/device, referral, domain, temporal</td>
</tr>
<tr>
<td>satisfaction indicators</td>
<td>troubleshooting email lists such as Libit-I, gateway-I, count-it, 404 page not found errors, turnaways, failed OpenURLs</td>
</tr>
<tr>
<td>work practice</td>
<td>this category refers to qualitative studies, such as surveys and interviews that tell us about the use of resources by the CU community, whether or not they are CUL resources</td>
</tr>
</tbody>
</table>

The full data set we collected for our inventory is included in two forms, both in the appendix. The complete set of raw data is included in the appendix as “Data sources inventory.” This appendix is long and boring, but to valuable readers who need the details of our usage data landscape. The second form in which we present the inventory is a spreadsheet that attempts to boil the contents into a more digestible form of yes or no responses to the four criteria described in the table above. There are many grey areas in trying to answer yes or no to the questions these criteria ask, and frankly, some of the responses in this spreadsheet could fit into either category. Our is for it to be used as a high level heuristic tool. We also include information about where usage data are being collected (local or external), which is critical to understanding both the
scope of our privacy challenge and the potential to mine the data to better understand our users.

4. Data classes and use cases

Cost per use: what is possible and how much effort is required?

Licensed ebooks

Use case

How often are ebooks used by CUL within a given time frame, and at what cost.

What is possible today?

Cost per use for individual e-books and e-journals may or may not be possible to determine, depending on how the CUL has purchased the materials. Many titles are purchased as part of a larger package, so we may not have a one-to-one payment; this may void the concept of cost per use for a specific title. It may, however, be possible to compute average cost/use for the entire package.

Cost per use of individual titles which are not part of a larger package are straightforward to determine by combining Voyager acquisitions and vendor supplied usage data. However, the unit of use for ebooks is less standardized than it is for ejournals. Some vendors report use at the title level, others at the "section" level, which is subject to variable interpretation. The size of a typical section may vary substantially from one title to the next, and the size of the average section may vary by publishing practice within different disciplines. Page limits may overrule section access. Some vendors report use at the individual page level, even in reports designed for section level reporting. All of these inconsistencies make the comparison of cost/use for ebooks less reliable than for ejournals.

Effort: Amount of effort and effective data varies depending on how the titles are purchased, and how use data is reported by the vendor. Low effort, for ebooks which are purchased individually, and the usage data is reported out in a standard format. Fairly low effort, too, to calculate the average cost of ebooks within a package, but the data may be less effective since it is giving an average cost per use for all of the titles within the package. Higher effort is required to glean cost per use data for ebooks which report out non-standard usage data. Effort increases greatly for packages if we try to use list prices to weight the average cost per use data to reflect that some books cost more than others.
What is possible with additional data, CUL human resources or outside services?

Unknown

Licensed ejournals

Use case

Compare the "value" of different titles through cost/use calculations. Use these calculations to guide cancellations (where allowed) or, in some cases, swaps between individually subscribed titles (with archival rights) and larger packages (without archival rights).

What is possible today?

Ejournals for which we pay an annual subscription fee permit computation of a yearly cost/use, assuming we have access to reliable use data (generally in the form of a COUNTER JR1 report). The unit of "use" for ejournals is fairly standardized to downloads at the article level. When we pay a one-time fee for perpetual access to a group of titles (say for a journal backfile), average cost/use can be calculated for the entire package, but the data isn't actionable. See also entry for ebooks.

Effort: Amount of effort and effective data varies depending on how the titles are purchased, and how use data is reported by the vendor. Low effort, for ejournals which are purchased individually, and the usage data is reported out in a standard format. Fairly low effort, too, to calculate the average cost of ejournals within a package, but the data may be less effective since it is giving an average cost per use for all of the titles within the package. Higher effort to glean cost per use data for ejournals which report out non-standard usage data. Effort increases greatly for packages if we try to use list prices to weight the average cost per use data to reflect that some books cost more than others.

What is possible with additional data, CUL human resources or outside services?

The large ejournal packages are quite complex to work with because they often consist of numerous mini-bundles, either optional or mandatory, wherein we pay a fixed amount for subscriptions to several different titles, making cost/use at the individual title impossible to compute.

However, despite limitations imposed by single line billing and minimal cancellation
provisions, the effort may be warranted in some cases where we have the flexibility to move titles between packages where we have the option to swap between packages of individually subscribed titles (with archival rights) and larger fixed cost packages (without archival rights).

There are several commercial services available to help manage use data from COUNTER reports, and to perform cost/use calculations. These include ScholarlyStats and SelectionSupport from Swets, 360 Usage Statistics from Serials Solutions, EBSCONet Usage Consolidation from EBSCO, and Ex Libris’ UStat. We used to subscribe to ScholarlyStats, but found it not to be cost-effective. It’s been a while since we looked into all of the available services and evaluated them.

**Licensed databases**

**Use case**

Analyze the use and/or value of licensed databases. May be used to compare similar databases to see which offers best combination of use, price, and value.

**What is possible today?**

Since databases are generally purchased as discrete resources (one subscription payment per database), combining the Voyager acquisitions data and vendor supplied usage data should be straightforward. Database use is often reported in the form of searches and/or sessions. Neither necessarily correlates very well with the value received by the user.

**Effort:** low effort for databases which use standard usage reporting; higher effort if the usage data is reported in a non-standard format.

**What is possible with additional data, CUL human resources or outside services?**

Counts of successful click-throughs to full-text resources, if available, could be used to enhance the usefulness of cost/use ratios for databases.
Print

Use case

Cost/use for print is most commonly calculated at the publisher level. It can be used to decide which publishers to put on approval and which to have selectors evaluate for purchase, or to otherwise fine tune approval plan profiles. It can also be used to judge the return on investment based on other characteristics, such as language of publication.

Also to support buy or borrow decisions, based on cost/transaction of interlibrary borrowing.

What is possible today?

We measure use of print materials in two ways, via circulation and browsing. Circulation is much more clearly defined, and more consistently measured over time and among different CUL units. Therefore, most analysis looks at circulation only. We retain circulation data for all print materials indefinitely (in some cases, longer than we keep the items themselves). We have total historical circulation data going back to about 1989. We have item level circulation data starting from mid-2000.

It is rare to compute cost/use of print books for individual titles. Generally, it is done at the publisher level. The assumption is that there is some consistency in the quality and popularity of publishing coming from each publisher, and that the publisher will continue to produce content of roughly the same quality. This may or may not be true, and we should probably compute standard deviation as well as average to distinguish publishers where the uniformity assumption is reasonable from those where it isn’t. We should also check our assumptions from time to time, since publisher output can change for the better or worse.

Other characteristics can be used to evaluate cost/use for print materials, including language of publication and subject.

Buy or borrow: The calculus of buy or borrow decisions involves quantification of many different expenses. When we buy, we generally know how much we paid the publisher or vendor for the item, plus shipping charges, but that’s only the beginning of the calculation. We must then add the cost of the staff that process the acquisition transaction, the staff that receive the shipment, that process the item through cataloging, and who prepare it for shelving. Then there are the costs associated with storing it, maintaining it, and circulating it over time. We also need to place value on the convenience for the patron. If a patron receives an ILL item and has only two weeks to use it, how does this compare to the convenience of a 1 year loan period? None of these costs are easy to estimate on a per item basis.
Borrow cost calculation are similarly complex. Expenses are incurred for both the borrower and the lender. For Borrow Direct, we ship multiple items to each partner each day, so shipping costs per item can only be approximated. There are no long-term costs once the item is returned, but also no equity (something like renting vs home ownership). If we borrow the same book multiple times, but don't keep track of that fact, then the buy or borrow comparison isn't very useful.

Effort (buy vs borrow): High effort, if all of the relevant factors are considered. Lower effort if only the most significant expenses for each option are considered.

What is possible with additional data, CUL human resources or outside services?

Individual cost of print books is stored in Voyager only for firm orders. For approval plan items, only the aggregate cost of each shipment is recorded. If we need to work with the cost of approval plan materials, we need data from YBP, Harrassowitz, etc. The data is available, upon special request in some cases.

For buy or borrow decisions, there are data sources, such as UPS WorldShip (the software used in the shipping department) that can help with the cost comparisons. Getting a true picture of borrow costs requires working with our ILL partners so the full cost of each transaction can be estimated.

**CUL hosted digital content**

To our knowledge, there is no effort currently underway within CUL to quantify the cost per use for our CUL hosted websites and digital content. The main problem here is to determine "cost" - are we to count salaries for people working on maintaining these sites, digitization costs, etc.? 

Depending on the site being analyzed, if budget records are available for how much the site cost to build and how much it costs to maintain, the effort to calculate the cost per use would be low. For some systems, such as digital collections, nearly all of the use is from non-Cornell users. We could take this important distinction into account if the logs for the site are in the CUL Logs system. The business model for arXiv is a good example of what can emerge from an analysis that is informed by usage demographics.
## Patron satisfaction

Table: Patron satisfaction: to what extent can we proactively identify frustrating service points using the data we have available?

<table>
<thead>
<tr>
<th>error type</th>
<th>use case</th>
<th>what is possible?</th>
<th>effort (low, medium, high, not possible)</th>
</tr>
</thead>
<tbody>
<tr>
<td>page not found on CUL website</td>
<td>Identify broken or missing content before patrons encounter it</td>
<td>All of our own hosted websites have error reports that contain the page not found errors.</td>
<td>low</td>
</tr>
<tr>
<td>page not found on licensed resource</td>
<td>Identify broken or missing content before patrons encounter it</td>
<td>Problems of this sort sometimes manifest themselves as subscription problems. Often, the problem is that the publication has migrated to a new platform, either because the publication was sold to a new vendor, or because the publisher has migrated to new platform.</td>
<td>Not possible for page not found errors. The first problem report of subscription problems are often unavoidable, but, if it is determined that related titles are affected, we can take action to avoid further problem. Platform changes can be handled pro-actively, if the publisher announces this information in a timely manner. Additional training (currently underway) needed to make sure that several staff understands what steps need to be taken when resources migrate to new platforms.</td>
</tr>
<tr>
<td>ebook/ejournal/database turnaways</td>
<td>Turnaways typically are not an issue for ejournals or databases, but they a very frustrating to our patrons trying to use our ebooks.</td>
<td>Our ebook and database vendors provide turnaway reports. The level of detail provided varies.</td>
<td>Effectively noting simultaneous use restrictions, particularly for ebooks, would help to minimize turnaways for class readings. CUL has made efforts, already, to enhance the availability of this information, and efforts continue.</td>
</tr>
<tr>
<td>ILL unfulfilled</td>
<td>Each month dozens or</td>
<td>CUL has retained</td>
<td>Olin ILS and LTS</td>
</tr>
<tr>
<td>error type</td>
<td>use case</td>
<td>what is possible?</td>
<td>effort (low, medium, high, not possible)</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>even hundreds of ILL requests go unfilled because the requested item is unavailable from other libraries and they do not fit the criteria for our buy or borrow program. From the perspective of our patrons, these amount to service failures.</td>
<td>complete records for all ILL transactions since 2000. We also have records of unfilled Borrow Direct requests.</td>
<td>acquisitions are currently running a pilot with a local vendor called Busca to reduce the number of unfulfilled ILL requests.</td>
</tr>
<tr>
<td>OpenURL failures</td>
<td>In 2012 some 516,000 requests were sent to our OpenURL link resolver. How many of these failed?</td>
<td>Technical Services samples for errors periodically to reduce the incidence of errors.</td>
<td>High. Analyzing OpenURL requests for errors is time consuming. NISO IOTA and KBART are both initiatives designed to address OpenURL quality at the network level.</td>
</tr>
<tr>
<td>Various</td>
<td>Identify non-working e-resources, or other barriers to access such as simultaneous use limits, browser/resource incompatibility issues, off-campus access problems, and others (which may be too numerous to list here).</td>
<td>We get reports of patron frustration in accessing resources through lists such as erlm-l, libgateway-l, and libit-l. Some complaints that come in via reference staff are documented in the Count-It system.</td>
<td>Medium to high effort to gather all of the disparate emails, and then to harvest the data included within the email. Also, may conflict with confidentiality, as we have a record of the patrons and what they are searching for. Low effort to gather data about a single resource.</td>
</tr>
</tbody>
</table>

**How users find and consume CUL content**

This group of metrics relates to the third question in the task force charge, "How do we tell we are making the right types of investment in digital services in support of electronic and other resources?" Examples include Voyager classic catalog (whether with a WebVoyage or Blacklight user interface), the Millennium e-Resource Management System, WorldCat Local, Summon, the Gateway’s Database Names and E-journal title services, and web sites for unit libraries that have been "virtualized." There are alternate means for finding relevant materials (search engines, Google.
Scholar, web browser bookmarks, etc.) users can bypass our own discovery systems and access licensed content directly from vendor sites. The question is, can we use data produced by these systems and other tools that monitor them to determine their effectiveness and the return on investment we get from developing, purchasing, and maintaining them?

We do have some data available now that can help us to understand how users are navigating our systems. This report, for example, shows the most heavily utilized CUL discovery interfaces when they clicked to access our licensed electronic resources on January 29, 2013. It was created using the CheckIP log that is available in the CUL Logs system.

<table>
<thead>
<tr>
<th>Referral URL</th>
<th>Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link resolver</td>
<td>1015</td>
</tr>
<tr>
<td>Find ejournals/databases</td>
<td>967</td>
</tr>
<tr>
<td>cornell.worldcat.org</td>
<td>419</td>
</tr>
<tr>
<td><a href="http://www.library.cornell.edu">www.library.cornell.edu</a></td>
<td>409</td>
</tr>
<tr>
<td>johnson.library.cornell.edu</td>
<td>247</td>
</tr>
<tr>
<td>guides.library.cornell.edu</td>
<td>192</td>
</tr>
<tr>
<td>olinuris.library.cornell.edu</td>
<td>51</td>
</tr>
<tr>
<td>catalog.library.cornell.edu</td>
<td>39</td>
</tr>
<tr>
<td>web.vet.cornell.edu</td>
<td>34</td>
</tr>
<tr>
<td><a href="http://www.vet.cornell.edu">www.vet.cornell.edu</a></td>
<td>30</td>
</tr>
<tr>
<td>engineering.library.cornell.edu</td>
<td>25</td>
</tr>
<tr>
<td><a href="http://www.lawschool.cornell.edu">www.lawschool.cornell.edu</a></td>
<td>21</td>
</tr>
<tr>
<td>mannlib.cornell.edu</td>
<td>21</td>
</tr>
<tr>
<td>alumni.library.cornell.edu</td>
<td>18</td>
</tr>
<tr>
<td>physicalsciences.library.cornell.edu</td>
<td>17</td>
</tr>
<tr>
<td>music.library.cornell.edu</td>
<td>16</td>
</tr>
<tr>
<td><a href="http://www.hotelschool.cornell.edu">www.hotelschool.cornell.edu</a></td>
<td>14</td>
</tr>
<tr>
<td>africana.library.cornell.edu</td>
<td>10</td>
</tr>
</tbody>
</table>

The following table presents a broad overview of the formats that could be analyzed.

Table: Overview of how users find and consume CUL licensed and CUL created or hosted digital content

<table>
<thead>
<tr>
<th>format</th>
<th>use case</th>
<th>what is possible</th>
<th>effort (low, medium, high)</th>
</tr>
</thead>
<tbody>
<tr>
<td>licensed ebooks</td>
<td>Determine patron use of catalog links to PDA titles</td>
<td>depends on availability of data from vendors</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>Determine effectiveness of our systems as gateways to licensed ebook content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>licensed ejournals</td>
<td>Determine effectiveness of our systems as gateways to licensed ejournal content</td>
<td>depends on availability of data from vendors</td>
<td>unknown</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>licensed databases</th>
<th>Determine effectiveness of our systems as gateways to licensed database content</th>
<th>depends on availability of data from vendors</th>
<th>unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>print</td>
<td>Determine effectiveness of our systems as gateways to owned print materials</td>
<td>Not traceable in most cases, unless patron requests office delivery, library-to-library transfer, or places a hold or recall</td>
<td>medium</td>
</tr>
<tr>
<td></td>
<td>Determine effectiveness of our systems as gateways to BorrowDirect print materials</td>
<td>fairly easy to track due to limited avenues for requesting BorrowDirect materials</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ILL requests made via OpenURL links usually include the source of the referral.</td>
<td></td>
</tr>
<tr>
<td>CUL hosted digital content</td>
<td>Determine effectiveness of our systems as gateways to our hosted digital content</td>
<td>For content hosted on our servers, we are able to track referrals using LOGS and AWStats</td>
<td>Low</td>
</tr>
</tbody>
</table>

## 5. Conclusion

There are two main themes that come out of this inventory.

(1) We need managerial oversight within CUL for the usage data that is being collected about our users. The person or people responsible will need authority and a coherent vision for how usage data should be used to improve services and help the Library reach its goals. Penn Library and its Data Farm is an example of what can be accomplished by an academic library with sustained attention and oversight over an extended period of time.

(2) Is it time to reconsider some of our long-held beliefs about patron privacy? Through the services CUL offers, gigabytes of usage data about our patrons are being collected and stored. As it stands, our policies about data retention are out of sync with our practices. We can either edit our policies to match our practices, or change our practices so they align with our policies.
6. Bibliography


Fang, Wei. 2007. “Using Google Analytics for Improving Library Website Content and Design: A Case Study.” *Library Philosophy and Practice*.


O'Rourke, JoAnne McFarland. 2012, "Disclosure Risk Assessment and Mitigation, An Overview," ICPSR (Interuniversity Consortium for Political and Social Research)
webinar.  
http://www.youtube.com/watch?v=2dw8_AA-a0s&list=UU4__DY4PxAg4Fispbubd-Pw&index=1&feature=plpp_video


sarahwordsworth. 2013.


“The Synthesis Project: Activity Data.”


Schwartz, Ray. 2011. “Application of EZProxy Logs, Voyager’s Patron Database, MySQL, and ColdFusion for Comparative Usage Analysis of Library Resources” presented at the ENUG Conference, October 27, University of Bridgeport, CT.


7. Appendices

Appendix: Overview of data sources
Appendix: Data sources inventory