Sampling the Umich.edu Domain

Jared Lyle

School of Information, University of Michigan
234 West Hall, 550 East University Avenue
Ann Arbor, MI 48109-1092, USA
lyle@umich.edu

Abstract. Using purposive, systematic and random sampling methodologies applied in the pre-Web archiving world, this case study tests the application of these same methodologies to web archiving of an entire domain: Umich.edu. While broad Web captures (e.g., Internet Archive) and selective downloads (e.g., Pandora) hold current legitimacy within web archiving, this study questions if sampling could improve the objectivity and overall merit of online material captured for present and future retrieval.

1 Introduction

> *This age of overabundant records and information, combined with increasing scarcity of resources, is forcing archivists to replace their essentially unplanned approach to archival preservation with a "systematic, planned, [and] documented process...."* [17]

Responding to the post-World War II deluge of records generated by both government and industry, archivists became destroyers in addition to their longstanding role as preservers [32, 41]. Now, thanks to computers and technology, archivists are once again afforded the opportunity to escape destroying in favor of preserving. Electronic records, whether "born digitally" or digitized, can be condensed, compressed, and stored at little cost and with little space. Technology is emigrated, emulated or integrated into new mediums when digital media deteriorates or becomes obsolete. Yet in this Age of Electronic Abundance, how are we to appraise as archivists: amass and hoard or cut and paste?

Using the sampling methodologies applied by the archival profession in the 1970's and 80's to paper documents, this case study applies sampling to the appraisal of current electronic records: Umich.edu web pages. That is, purposive, systematic and random samples are used to appraise the over four million Umich.edu related documents captured from a January 2004 Internet Archive web crawl. Through attempting to selectively cut and paste, this study tests sampling as a viable means of appraising electronic records for a lasting and useful present and future.

2 Appraising the Web

Appraisal, or the estimation of value, has been described as "the greatest professional challenge to the archivist" [11]. Indeed, it is the gatekeeper through which an archives lives and grows (and, sometimes, also shrinks). Appraisal allows for acquisitions that best match the scope and purpose of a collecting institution, while also ensuring a manageable and usable collection – one through which the user may navigate without becoming encumbered by extraneous material. Yet, how does one go about appraising, especially with respect to electronic records – archivists' "greatest challenge in decades" [19]?

2.1 Passive Appraisal

Early archival theorists and practitioners practiced passive appraisal of their collections. Muller, Feith and Fruin, of late 19th Century Dutch manual fame, believed that "the rules which govern the composition, the arrangement and the formation of an archival collection…cannot be fixed by the archivist in advance; he can only study the organism and ascertain the rules under which it is formed" [39]. Jenkinson, three decades later, also emphasized a hands-off approach to appraisal, considering the archivist as a "passive professional, receiving…records and guarding them for posterity" [3]. Although Jenkinson and the Dutch archivists practiced limited appraisal through primarily accepting materials "of an official transaction" [27] or from "an administrative body or one of its officials" [39], they ultimately deferred appraisal responsibility to the record creators and, instead, placed greater importance on concepts of unbroken custody, provenance and *respect des fonds*.

2.2 Active Appraisal

Schellenberg and Booms, however, pressed toward active appraisal by archivists. Working in the mid- to late-20th Century, they sought to balance the need to preserve with the reality that archival space is finite. Post-World War II bureaucracies along with advances in duplication technology had dramatically increased the potential materials to be stored within archival institutions, and the time was at hand for archivists to address the issue. "Not all [modern records] can be preserved," wrote Schellenberg [44], "…some of them have to be destroyed, and…a discriminating destruction of them is a service to scholarship." Likewise, Booms [5] noted that

> if our archives are not to degenerate into storehouses of antiquarian curiosities, then we must be serious in our efforts to overcome decisively the most widespread challenge of our profession: to reduce the growing quantity of documentation to the form of a documentary heritage that is of a *useable and storable* quality. [emphasis added]

For Schellenberg, active appraisal consisted in "moderating" value. While items have inherent, primary value "for the originating agency itself," he wrote, their secondary values "will exist long after they cease to be of current use" [44]. By appraising a record by its informational value, or "content of the records themselves"

[14], and its evidential value, or the evidence of "the functioning and organization of the…body that produced them" [44], archivists may retain records of highest value for future generations.

For Booms, active appraisal consisted in searching out those records that best reflect society. "Measuring the societal significance of the past facts by analyzing the value which their contemporaries attached to them should serve as the foundation for all archival efforts toward forming the documentary heritage" [5]. Through analyzing public opinion, archivists could then make appraisal value judgments.

While Schellenberg and Booms nudged the archivist into an assertive stance toward appraisal, their models leave opportunities for divergence. Most notable is their reliance upon subjective methodologies that rely heavily on human judgment. Does an item with high informational value always have greater value than an item with high evidentiary value, or vice versa? Can one really divine the values of society with impartiality? Must record value be deduced through searching individual records?

2.3 The Active and Objective Appraisal of Electronic Records

With the addition of electronic records to the fray of archival appraisal, archivists have been scrambling even harder to objectify the appraisal processes while also continuing to actively appraise. Central to this effort is the inclusion of appraisal at the creation phase of the record life cycle [9]. As Catherine Bailey [13] said,

> [Archivists] cannot wait until inactive electronic records are offered to them for appraisal, as they might have for paper records; too many computer records have vanished by then, and the documentation necessary for their proper appraisal has been lost, destroyed, or is hopelessly outdated. The sheer volatility of electronic records should be a powerful inducement for archivists to accept increased involvement in the scheduling process, beginning at the systems design stage.

Bringing the appraisal process to the beginning of the record creation ensures eventual records capture, while also providing for a more objective approach to record valuation by using input from both record creator and archivist. Metadata automatically generated with electronic records can more clearly demarcate value without the necessity of human judgment. But what of the electronic records on the World Wide Web, where information is not clearly demarcated according to function or purpose, and where information instantaneously appears and disappears? Is there an appraisal model that can create an unbiased and usable collection for the present and future?

3 Capturing the Web

As previously mentioned, the ephemeral nature and sheer volume of electronic records are warranting active and objective appraisal by archivists. E-mail messages per year within the U.S. federal government, for instance, were estimated at 36.5

billion in 1999, which number has likely doubled or quadrupled in the five years since [20]. The Internet is growing exponentially, with 233,101,481 total IP addresses in April 2004 that have been assigned a name compared with a *mere* 29.6 million addresses assigned in 1998 [30]. The mind-boggling 2003 report, "How Much Information" [36], only magnifies the evidence of information overload by illustrating that around 5,000,000,000,000,000,000 bytes of new information were produced in the year 2002 alone.

3.1 Return to Passive Appraisal

Perhaps caving in to these tidal waves of potential archival material – especially with reference to electronic records on the World Wide Web – archivists are beginning to turn back to hands-off, passive appraisal. That is, with costs of electronic storage consistently declining and automation enabling and simplifying large scale data capture, archivists can preserve the entirety of a collection without actively appraising for value. Data storage fees, while prohibitively expensive in 1992 when one gigabyte cost nearly $1,000, are plummeting, with estimates projecting the same amount of storage to be a mere two cents by 2010 [16]! By storing everything, archivists need not impute value or limit collections.

Internet Archive. The Internet Archive [26], for example, a digital archiving venture begun in 1996, attempts to aggressively capture the entire Web in two-month cycles (over 300 terabytes have been downloaded so far). Using robots (a.k.a. spiders) to crawl the Web for potential material, captured sites are available for retrieval through a public interface. The only appraisal evident in this process is the non-capture of sites with embedded messages asking for robots to leave the web site alone (i.e., not capture it).

Kulturarw3. The Kulturarw3 project, like the Internet Archive, broadly crawls the Web in search of material to capture. Although it attempts to only capture Swedish web pages, the vast content of the Web registered with addresses ending in '.se' are millions in number (at least 1,539,917 hosts alone, not counting web pages within each host [30]), notwithstanding the myriad Swedish sites on other top Web domains such as .com or .org. As such, the Kulturarw3 project attempts to "preserve everything [of Swedish content] with the aid of computer technology" [emphasis added] citing its justifications following the difficulty in establishing what will be of value to future researchers and what will not. To make selections from millions of web sites require enormous personnel efforts at high costs, whereas costs for data memory storage are decreasing at a rapid rate [42]. The most recent Kulturarw3 crawl "gathered…approx.30 million URL files extracted from 126.000 web sites of which two thirds were found filed under .se. The total quantity of data is ca [sic] 1350 Gbyte." Users can access this raw information on computers at Sweden's National Library, the Royal Library.

3.2 Active but Biased Appraisal

Other equally aggressive projects devoted to archiving the Web have been more selective, and have included a greater degree of appraisal. Yet, increased human subjectivity and bias of selection have come with such selectivity.

Pandora and MINERVA. Pandora, a project of the National Library of Australia, is "dedicated to the preservation of and long term access to Australian online electronic publications of national significance" [40]. Selection criteria include locating government publications, conference proceedings, e-journals, and sites of "current social or political interest." Pandora's archive size, as of March 26, 2004, was 18,898,293 files comprising 606 gigabytes. Another electronic archiving project involving some initial appraisal for selection includes the Minerva project of the American Library of Congress. Criteria for capturing pages include: "usefulness in serving the current or future informational needs of Congress and researchers, unique information provided, scholarly content, at risk of loss (due to ephemeral nature of Web sites), and currency of the information" [35]. Ultimately, sites are "determined by Recommending Officers in consultation with the MINERVA Team." The public may view the selectively captured sites on the Library of Congress's web page.

3.3 Limitations of Current Web Archiving Techniques

Of immense benefit to a world-wide audience, all four of the aforementioned projects uniquely attempt to capture and archive the expanding and inchoate Web. Unfortunately, their attempts are also yielding biased and subjective collections that do not take full advantage of the unique technological environment of the Web. The massive crawls are harnessing the brute force of inexpensive storage and automated collection and overlooking appraisal through tacitly deeming everything captured as holding potential future value. Likewise, selective crawls do not take full advantage of the powerful Web and rely too heavily on human judgment when appraising. Can one find a happy medium, where relevant material encompassing the entirety of the human experience is captured in manageable, yet noteworthy, portions?

4 Sampling the Web

Introduced in 1957 to archival literature by Paul Lewinson of the U.S. National Archives, archival sampling is an appraisal method used to select "some part of a body of homogenous records, so that some aspect of [an organization's] work or the information received or developed by [the organization] may be represented or illustrated thereby" [34]. Chronicled primarily in the 1970's and 80's, sampling was documented as a viable means of logically or systematically decreasing the overall bulk of paper records to a manageable size, while still preserving the overall representation of that same sample. Although archival literature has since moved

onward, sampling might well serve as a viable alternative to appraising electronic records on the World Wide Web.

Some of the most widely reported appraisal cases within the archival world have involved sampling. The appraisal of the records of the U.S. Federal Bureau of Investigation through sampling in the early 1980's, for instance, was described by one prominent archivist as the "most important records appraisal ever undertaken in this country" [6], and involved over 500,000 cubic feet of records spread over seventy locations, with seventeen archivists – "the number normally assigned to appraise the records of the entire federal government" [6] – working the project [51]. Similar large-scale projects transpired in appraisals of the records of the Massachusetts courts [21], the Quebec courts [2], and the U.S. Department of Justice litigation case files [50].

Comparatively smaller-scale appraisals have also successfully employed sampling. Eleanor McKay [37], in "Random Sampling Techniques: A Method of Reducing Large, Homogenous Series in Congressional Papers," illustrated how sampling could reduce the bulk of U.S. Congressional papers, where, for instance, one U.S. Congressman cached more than 400 cartons over a decade in office, to a manageable and usable collection. Likeminded smaller-scale projects included sampling timber company records [47] and silver-lead mining records [12].

4.1 Sampling Criteria

How does one go about sampling? Besides determining that "a series is bulky, homogeneous, and repetitive" [37], and that space, budget and staff resources warrant a significant weeding-out of a particular collection [46], an archivist must also discern what is beneficial to the potential users. Records consistently high in informational content, for example, would not be good candidates for sampling, where a sample would only decrease the value of the collection (e.g., Presidential daily briefings before and after September 11th). Likewise, records of low information value (e.g., celebrity fan mail) also might not be of worth to sample, as they are difficult to arrange in order to first run a sample. Margery Sly [46] provides a practical and salient framework of "Archival Considerations" when contemplating sampling:

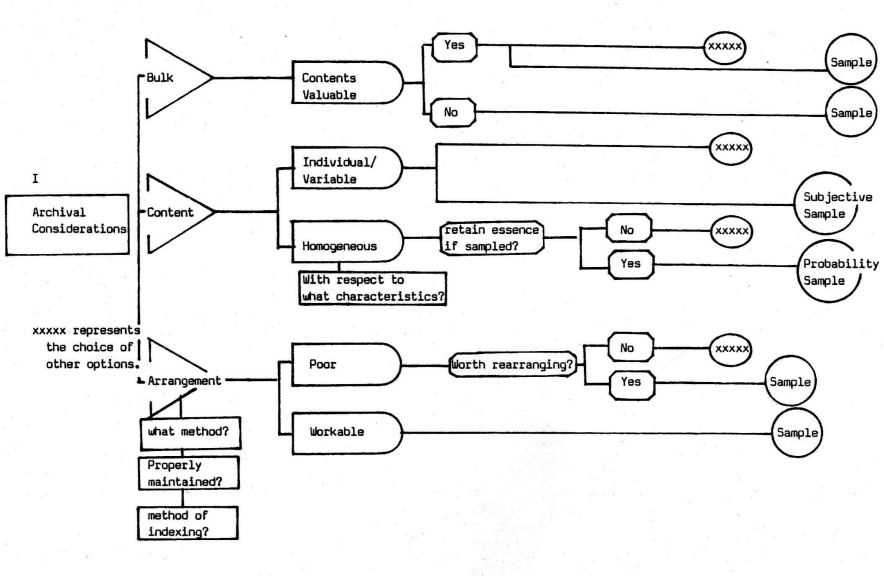

Fig. 1. Archival Considerations.

Knowing one's collection is worth sampling, an archivist can then determine which type of sample will best produce a representative outcome. Three of the primary sampling methods within archival appraisal include purposive, systematic and random sampling.

4.2 Purposive Sampling

Purposive sampling, "or qualitative sampling on a pre-determined pattern or basis" [24], is the most basic form of sampling, and, consequently, is the easiest of samples to perform. The Dutch Ministry of Justice, for example, used a purposive sample of its one million personal files to select "files which are formed in deviation from the rules. Here, only the exceptions (around 3 percent of the total) to the common practice are preserved" [1]. Unfortunately, purposive sampling is also highly prone to human bias and selection error. Through dealing subjectively with the appraisal process, one's samples are skewed toward personal feelings and cannot be taken as representative of the whole.

4.3 Systematic Sampling

Systematic sampling, or "selection based on a physical characteristic of the records or filing scheme without regard to the substantive information in the selected files" [11], is less prone to bias than purposive sampling, yet still prone to error, as the order of collections may be unduly weighted in particular areas so as to skew the resulting samples (e.g., consider systematically selecting files beginning with A if the

population sampled is Chinese). Common forms of systematic sampling include: alphabetical, numerical and chronological samples (i.e., taking every n^{th} record in order of alphabet, number or date), and physical samples (e.g., the "fat file method," where files are selected according to thickness) [1]. Interestingly, while systematic samples technically remain biased, they can be excellent tools in searching out archival value. The Massachusetts Superior Court appraisal, for example, found case files to increase in archival value with size [21]. Canadian research found similar results [43]. Systematic samples, such as the Finnish attempt to sample all National Land Settlement Board records from people born on the 18^{th} and 28^{th} days of each month [33], also amply store a representative picture of society.

4.4 Random Sampling

Random, or statistical, sampling is, in theory, the most objective of the three main sampling methodologies. It allows for every record within a sample equal opportunity to be chosen. Whether using a purely random sampling or a systematic random sample (i.e., begin with a random number and then choose every n^{th} case), or even a stratified random sample (i.e., randomly choosing within varying sectors of a population), random sampling ensures accuracy, validity and comprehensiveness [10]. Users of the sample are given "a clearer idea of what they are using and how much faith to put in it" [48]. Yet, with such objectivity comes work; should an archivist misjudge a collection's characteristics, a supposed random sample would turn out to not be representative. Random sampling can also be labor intensive, especially when preparing paper files for sampling (i.e., through numbering and pre-arranging files) [31].

4.5 Mixed-Mode Sampling

An additional method of sampling is also available. This involves combining the various forms of sampling to provide for an amplified result. For example, knowing that random sampling provides a representative sample, an archivist might use a systematic skimming of "fat" files to also include records of particular high value. In these instances, nonetheless, where objective samples are supplemented by subjective samples, the subjective sample "should be selected after the [objective,] statistical sample is developed to ensure statistical validity" [7].

5 Sampling the Umich.edu Domain

"It is surprising to many that archival theory, developed in a paper world, appears to be logically valid in the electronic age" [49]. Building upon the sampling theories applied to the paper world, this case study looks at how well similar models of sampling apply to the World Wide Web. It is hoped that sampling will allow archivists to objectively harness the powerful Web when appraising, so to help weed out bulk and redundancy while also creating a usable and manageable collection. For,

as the East German Angelika Menn-Haritz [14] said, "the aim of archival appraisal, as regards traditional material also, is not to reduce quantities, but to make archives eloquent and to facilitate research."

5.1 Crawling the Umich.edu Domain

Using Heritrix, an "open-source, extensible, web-scale, archival-quality web crawler project" [25], the Internet Archive performed three narrow crawls of the Umich.edu domain – beginning at www.umich.edu – on January 8-9th, January 14-15th, and January 19-20th [38]. These crawls were coordinated in conjunction with an archival practicum course taught by Dr. Margaret Hedstrom, Associate Professor of Information at the University of Michigan School of Information, and were separate from the general crawls the Internet Archive regularly performs [18].

The Umich.edu domain services the entire University of Michigan community, which includes over 51,000 students and 5,600 faculty – as well as numerous staff – spread across three distinct campuses. Although usage of the Umich.edu domain is centrally administered, serving of pages to the Web is dependent upon many satellite servers. These servers run independent of the server controlling the top-level Umich.edu pages. The University Health System, School of Engineering, and School of Information, for example, all serve their Umich.edu domain pages from their own independent servers. The Institute for Social Research, a university affiliated organization, also serves its Umich.edu associated pages from its own server(s). Thus, while Umich.edu domain administrators heavily control top-level pages, lower level material, such as departmental personal pages, are informally managed by whoever serves the pages to the Web [23].

Two hundred and twenty gigabytes were captured in the Umich.edu crawls, which is similar to two library floors of printed academic journals, 1,760 boxes, or 3,520 linear feet of storage [36, 16]. The majority of the captured information – generically termed documents [25] – was saved in large aggregate files [8], although smaller crawl logs for each of the crawls were also available.

The crawl logs contained particularly useful metadata pertaining to each document captured, which greatly assisted with the analysis after all three logs were 'flattened' into one cumulative SPSS file. These metadata included [25, 38]:

- Timestamp (e.g., '20040721232940438').
- Status code (e.g., '200') [15].
- Size, in bytes (e.g., '413').
- URI of the document downloaded (e.g., www.umich.edu).
- Server returned MIME format (e.g., 'text/html').
- Hop-types, showing how each document was reached by the crawler (e.g., 'R' (redirect)).
- Precurser URI in the crawl hop-type chain (e.g., 'www.umich.edu' for the URI 'www.umich.edu/~info').

It was from these crawl logs that the following sampling methodologies were applied.

5.2 Crawl Population

Exactly 4,064,170 Umich.edu related document URI's were captured in the crawl log. Unfortunately, not all of these documents were retrievable at the time of the crawl. Roughly 87 percent (3,525,633) had operative status codes. Using the World Wide Web Consortium's "Status Code Definitions" as a reference [15], operative status codes were defined as those codes showing a document's URI to be either "Successful" (i.e., 2xx) or a "Redirection" (i.e., 3xx). URI's with inoperative codes included documents that could not be retrieved by the crawler, primarily due to bad links or off-line pages.

Within the roughly 3.5 million operative Umich.edu web documents captured, an estimated 1,960,863 were duplicates. This was to be expected with three separate web crawls of the Umich.edu domain. Thus, only 39 percent (1,564,770) of the total captured document URI's were shown to be both operative and unduplicated.

Table 1. Crawl Population.

Status	#	%
All Umich.edu web documents originally captured in crawls.	4,064,170	100%
Umich.edu web documents with operative URI status codes.	3,525,633	87%
Unduplicated and operative Umich.edu web documents.	1,564,770	39%

5.3 Sampling Results

Using the roughly 1.5 million operative and unduplicated Umich.edu web documents, purposive, systematic, and random samples were conducted to determine if the resulting samples could produce a manageable, unbiased appraisal of the existing Web.

Purposive Sampling. Testing the validity of purposive sampling, crawl logs were scanned in search of Umich.edu documents with high informational and historical content. Sites, such as those belonging to university departments (e.g., http://www.law.umich.edu), as well as sites belonging to student organizations (e.g., http://www.engin.umich.edu/solarcar), were culled from the logs using human selection.

Unfortunately, purposive sampling as a stand-alone sampling methodology did not provide for more unbiased or manageable appraisal of web sites. Perhaps this is an obvious result, as sites were captured according to the human selector's criteria. Although it could be said that through online search capabilities one can personally scan information at a faster rate, such appraisal may even leave out more from the intended collection due to the limited scope of a predetermined search (e.g., using the search command within a web browser as opposed to scrolling down a document).

Systematic Sampling. Two forms of systematic sampling were conducted using the captured Umich.edu documents. First, noting that previous American studies used "fat" files as one criterion of selection, the Umich.edu log documents were sorted according to size. Documents with huge sizes, according to the theory, would be of greatest significance.

Interestingly, systematically selecting according to size, or "fat" documents, did not hold up within the electronic environment. Many of the huge documents within the captured Umich.edu domain were simply images and applications. Text files, while seemingly just as important as images and applications, tended to have smaller sizes, and would consequently slip under the systematic "fat" file radar if such a sampling method were to be used.

The second form of systematic sampling tested was selecting Umich.edu documents with short URI's. In theory, shorter URI's would be higher-level pages with higher informational content. For example, www.housing.umich.edu is the home page for the University of Michigan's housing department. It is the central page from which all other housing pages originate (at least, in theory).

The results indeed displayed many of the high-level pages within the Umich.edu domain, which can be useful considering the university's large number of servers maintained by numerous interrelated yet unrelated departments. One can envision the population and general make-up of the domain sites. This sample, however, was also heavily biased toward showing official Umich.edu pages – especially pages relating to university departments and programs. Personal pages, for example, were completely lost in this form of systematic sample.

Random Sampling. Supposedly the least biased of the methods of archival sampling, two forms of random sampling were conducted with the Umich.edu crawl logs. The first form, running a systematic random sample, consisted of ordering all captured document URI's alphabetically and then choosing every nth URI (in this case, every 985th URI). While the method was certainly systematic and deliberate, it also skipped over entire series of documents and didn't provide for much of a representative sample. Because URI's aren't meant to be arranged alphabetically, systematic random sampling by alphabet largely just biased the results.

Conversely, using a stratified random sample proved to be a success, and granted objectivity and freedom from personal bias. Noting that 60,190 of the Umich.edu documents were personal pages (i.e., as determined through containing the word "personal" in the URI), a simple random sample was performed on the singular stratum. The random sample effectively provided a less biased and more objective representation of the general Umich.edu personal pages. Because no known master list of personal pages exists within the university community, randomly sampling for personal pages provides both a wide spectrum and interesting range of documents.

6. Discussion

What is there to learn from the test samples from the Umich.edu crawls? Can these lessons be generalized to the greater archival community, especially those appraising electronic records and the World Wide Web?

Admittedly, sampling did not prove to be the answer of all answers to archival appraisal. It did, however, provide unique insight into potential future use and research. Sampling is an excellent means of gathering a "big picture" of record series. Beginning with 4 million Umich.edu web documents, sampling provided a generalized look at a huge population. It also was a method of finding the exemplary records within the population. Better yet, sampling, under the correct conditions, provided for greater objectivity and more thorough scouring of the records (e.g., stratified random sampling of personal records). As Hite and Linke [22] have stated,

> Archivists should use statistics to support their ideas….In the area of appraisal, if repositories compiled records of removal rates and analyzed their data, then found appropriate venues to disseminate their findings, it would be a step toward an empirical approach to appraisal.

Sampling is one such empirical approach to archival appraisal.

Yet, "sampling," says one writer [29], "while a valuable tool, is not a panacea for archivists." Sampling cannot magically uncover all records of lasting value within a series. It can, however, provide a relatively unbiased approach to selecting a representative portion of a larger collection. As the distinguished Margaret Norton [48] has said, "It is comparatively easy to select records of permanent value, relatively easy to decide on those of no value. The great bulk are borderline." With sampling, the borderline records are on an even footing with those of permanent and zero value.

Still, sampling remains challenging. Records that are continuing or open-ended are difficult to sample, as a representation from a population with vague parameters is sketchy, where results cannot be compared to a known universe [10]. Likewise, samples of samples cannot be deemed representative or significant. Samples of web crawls, for example, are more comparable to snowball sampling – a form of sampling that relies heavily upon relationships, which is not representative or unbiased. Longitudinal samples over time are also difficult to achieve, as truly random samples will not capture duplicate information over multiple surveys, especially when crawls do not capture the same information over and over.

Other dilemmas archivists encounter with sampling have included high costs and tedious amounts of time [28, 31]. Anything to do with statistics or numbers, likewise, sometimes scares off prospective samplers. These complexities, however, have more to do with the world of the 1970's and 80's, during which sampling was still relatively noticeable within archival literature. While Frank Boles stated in 1981 that "the most vexing procedural difficulties of simple random sampling involve the use of random number tables" [4], recent advances in statistical software, such as SPSS, all but vanquish the tedium of statistics to yesteryear. Simple drop-down menus and online help make generating random samples a relative breeze compared to the intricate random number tables of the past. Similar computer technology also reduces the time and tedium previously associated with sampling. No longer does the archivist need to number all ten thousand of her files before systematically choosing a

sample. Software, such as Microsoft Access, can autonumber records with the click of a button. Time spent manipulating and analyzing the Umich.edu crawl logs in SPSS, for example, was in hours, not days or weeks.

Perhaps the most ominous of sampling quandaries involves the inability of archivists to throw things out [45]. All archivists will likely have one sleepless night during a professional career regarding something they personally have removed from their collection. But with sampling, where records are appraised using a relatively non-human process (e.g., random sampling) by means of a computer, an archivist might feel extra anxious regarding his or her 'powerlessness' with respect to a random sampling of records. Still, archival appraisal remains a knotty task, with or without sampling. As Leonard Rapport [7] said, when appraising records there is "one immutable law: there are no perfect appraisals and the best appraisal is the one that does the least harm."

That said, Cook's [10] caution regarding sampling holds true today, especially with electronic records: "Sampling is a powerful tool, therefore, but should be used sparingly and only when all the conditions for statistical validity can be met and all other appraisal options have been considered first." Still, archivists should not fear to push the electronic envelope in this modern era. Sampling, under the proper conditions, can provide an alternative means of less biased and more manageable appraisal of archival collections. Although systematic and purposive samples applied to the Umich.edu documents proved inadequate, random sampling of personal documents showed promise. Perhaps similar samples to other sub-collections within the domain would show similar merit. Why save it all, especially when many documents on the Web are redundant? Why not take the initiative and sample?

Acknowledgements

Special thanks to Dr. Margaret Hedstrom for initiating the U-M crawl and critiquing an earlier draft, Dr. David Wallace for critiquing an earlier draft and presentation, Rachael Hu and Leslie Knoblauch for their practicum teamwork and ideas, and the Internet Archive – especially Gordon Mohr, Molly Davis and Brewster Kahle – for compiling and explaining the January crawls.

This work is licensed under a Creative Commons License.
http://creativecommons.org/licenses/by-nd/2.0/

References

1. Aerts, E.: The Use of Elementary Sampling in the Appraisal Process of Records. In: Black-Veldtrup, M., Dascher, O., Koppetsch, A. (eds.): Archive vor der Globalisierung? Veröffentlichungen der Staatlichen Archive des Landes Nordrhein-Westfalen, Düsseldorf (2001) 157-206
2. Archives National du Quebec : Report of the Interministerial Committee on Court Records. Archives National du Quebec, Montreal (1991)
3. Blouin, F.: Archivists, Mediation, and Constructs of Social Memory. Archival Issues 24 (1999) 101-112
4. Boles, F.: Sampling in Archives. The American Archivist 44 (1981) 125-130
5. Booms, H.: Society and the Formation of a Documentary Heritage: Issues in the Appraisal of Archival Sources. Archivaria 24 (1987) 69-107
6. Bradsher, J.G.: The FBI Records Appraisal. The Midwestern Archivist 13 (1988) 51-66
7. Bradsher, J.G., Ambacher, B.I.: Archival Sampling: A Method of Appraisal and A Means of Retention. Mid-Atlantic Regional Archives Conference Technical Leaflet Series, Number 8. (1992)
8. Burner, M., Kahle, B.: Arc File Format. (1996) http://www.archive.org/web/research/ArcFileFormat.php
9. Cook, M.: Archives and the Computer. Butterworths, London (1986)
10. Cook, T.: 'Many are called but few are chosen': Appraisal Guidelines for Sampling and Selecting Case Files. Archivaria 32 (1991) 25-50
11. Cook, T.: The Archival Appraisal of Records Containing Personal Information: A RAMP Study with Guidelines. UNESCO, Paris (1991)
12. Davis, R.C.: Getting the Lead Out: The Appraisal of Silver-Lead Mining Records at University of Idaho. The American Archivist 55 (1992) 454-463
13. Eastwood, T., Katuu, S., Killawee, J., Whyte, J.: Appraisal of Electronic Records: A Review of the Literature in English. (1999) http://archivi.beniculturali.it/Divisione_V/convenzioni/censimento/appraisal.html
14. Erlandsson, A.: Electronic Records Management: A Literature Review. International Council on Archives, Paris (1996) http://www.ica.org/biblio/litrev.txt
15. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext Transfer Protocol – HTTP/1.1: Status Code Definitions. (1999) http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html
16. Gilheany, S.: Projecting the Cost of Magnetic Disk Storage Over the Next 10 Years. (2004) www.archivebuilders.com/whitepapers/27006p.pdf
17. Ham, F.G.: Archival Choices: Managing the Historical Record in an Age of Abundance. American Archivist 47 (1984) 11-22
18. Hedstrom, M.: SI 692 practicum projects hand-out. (2004)
19. Hedstrom, M.: Understanding Electronic Incunabula: A Framework for Research on Electronic Records. American Archivist 54 (1991) 334-54
20. Henry, L.J.: Appraisal of Electronic Records: Traditional Principles Endure. In: Ambacher, B.I. (ed.): Thirty Years of Electronic Records. The Scarecrow Press, Inc., Lanham, Maryland (2003)
21. Hindus, M., Hammett, T.M., Hobson, B.M.: The Files of the Massachusetts Superior Court, 1859-1959. G.K. Hall and Company, Boston (1991)
22. Hite, R.W., Linke, D.J.: A Statistical Summary of Appraisal During Processing: A Case Study with Manuscript Collections. Archival Issues 17 (1992) 23-29
23. Hu, R., Knoblauch, L.: Personal correspondence with the author. (2004)
24. Hull, F.: The Use of Sampling Techniques on the Retention of Records: A RAMP Study with Guidelines. UNESCO, Paris (1981)

25. Internet Archive: Heritrix (2004) http://crawler.archive.org
26. Internet Archive: Internet Archive Home Page. (2004) http://www.archive.org
27. Jenkinson, H.: A Manual of Archive Administration. Clarendon Press, London (1922)
28. Kepley, D.R.: E-mail correspondence with the author. (2004)
29. Kepley, D.R.: Sampling in Archives: A Review. The American Archivist 47 (1984)
30. Kessler, J.: Internet in France, 2004. (2004) http://www.fyifrance5.com/Fyarch/fy040315.htm
31. Kolish, E.: Sampling Methodology and its Application: An Illustration of the Tension Between Theory and Practice. Archivaria 38 (1994) 61-73
32. Lamb, W.K.: The Fine Art of Destruction. In Hollaender, A.E.J. (ed.) for the Society of Archivists: Essays in Memory of Sir Hilary Jenkinson. Moore and Tillyer, Chichester, UK (1962) 50-56
33. Leppänen, M.: The Use of Sampling in the Appraisal and Disposal of Records. (date unknown) http://www.narc.fi/parnu/3.pdf
34. Lewinson, P. : Archival Sampling. The American Archivist 20 (1957) 291-312
35. Library of Congress: Collections Policy Statement: Web Site Capture and Archiving. (2004) http://lcweb.loc.gov/acq/devpol/webarchive.html
36. Lyman, P., Varian, H.: How Much Information 2003? (2003) http://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm
37. McKay, E.: Random Sampling Techniques: A Method of Reducing Large, Homogenous Series in Congressional Papers. The American Archivist 41 (1978) 281-289
38. Mohr, G.: E-mail correspondence with the author. (2004)
39. Muller, S., Feith, J.A., Fruin, R.: Manual for the Arrangement and Description of Archives. H.W. Wilson, New York (1968)
40. National Library of Australia and Partners: Pandora Archive: Preserving and Accessing Networked Documentary Resources of Australia. (2004) http://pandora.nla.gov.au/index.html
41. Peace, N.E.: Deciding What to Save: Fifty Years of Theory and Practice. In: Peace, N.E. (ed.): Archival Choices: Managing the Historical Record in an Age of Abundance. D.C. Heath and Company, Lexington, MA (1984) 1-18
42. Royal Library, National Library of Sweden: Kulturarw^3. (2004) http://www.kb.se/kw3/ENG/Description.htm
43. Scheinberg, E.: Case File Theory: Does It Work In Practice? Archivaria 38 (1994) 45-60
44. Schellenberg, T.R.: Modern Archives: Principles and Techniques. The Society of American Archivists, Chicago (1956)
45. Sly, M.N.: Personal correspondence with the author. (2004)
46. Sly, M.N.: Sampling in an Archival Framework: Mathoms and Manuscripts. Provenance 5 (1987) 55-75
47. Steck, L., Blouin, F.: Hannah Lay & Company: Sampling Records of a Century of Timbering in Michigan. The American Archivist 39 (1976) 15-20
48. Stoddart, M.: Sampling – The Much Maligned Archival Choice – A New Zealand View. Archifacts: Bulletin of the Archives and Records Association of New Zealand October (1989) 19-32
49. Turner, J.: Theoretical Dialectics: A Commentary on Sampling Methodology and its Application. Archivaria 38 (1994) 74-78
50. United States, National Archives and Records Administration, Office of Records Administration: Appraisal of Department of Justice Case Files: Final Report. National Archives and Records Administration, Washington, D.C. (1989)
51. U.S. National Archives and Records Service: Appraisal of the Records of the Federal Bureau of Investigation: A Report to Hon. Harold T. Greene, U.S. District Court for the District of Columbia. U.S. National Archives and Records Service, Washington, D.C. (1981)