IBIBLIOMETRICS:

MEASURING TEN YEARS OF FREE-RANGE ACADEMIC WEB PUBLISHING

By:
Donald Sizemore, II

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

March, 2003

Approved by:

_____

Advisor

Donald Sizemore. ibibliometrics: Measuring Ten Years of Free-Range Academic Web Publishing. A Master's paper for the M.S. in I.S. degree. March, 2003. 40 pages. Advisor: Gary Marchionini

This research is a bibliometric study of descriptive information on the individual sites comprising the *ibiblio.org* academic web server. It hopes to acquire a snapshot of the general content, age, and sizes of the collections as a whole, and to look for any patterns or similarities among the publishing habits of its contributors.

Initial measurements of dates, sizes and genres of ibiblio collections and other web directories revealed that one of the Internet's oldest and largest public archives suffers little of the stagnation and "collection rot" that would be expected of such an organization. Its content slants towards the *Arts and Entertainment, Technology,* and *History* UDC categories, and that most of its physical collections web space is consumed by multimedia content.

Headings:

Information Systems

Digital Libraries

Bibliometrics

Publishing Trends

Acknowledgements:

First and foremost I would like to thank Paul Jones for his constant advice and guidance, invaluable insight, lethal wit, and five years of gracious employment. Thanks to my advisor, Gary Marchionini, for his advice and suggestions through the entire project. Thanks to my co-workers: Jonathan Magid for years of explanations and guidance, John Reuning for his technical advice, Adriane Boyd for reality checks, Miles Efron for his reviews and reassurance, and Fred Stutzman and Serena Fenton for their encouragement and humor. Finally, thanks to the staff of the Daily Grind Coffeeshop for their comfort and their caffeine, which fueled most of this paper.

TABLE OF CONTENTS:

LIST OF TABLES:

LIST OF FIGURES:

I. Introduction

In his *New Community Networks,* Doug Schuler observes that “the old or 'traditional' community was often exclusive, inflexible, isolated, unchanging, monolithic, and homogeneous” (Schuler 1996, p. 9). Early Internet technologies, developed to serve military interests and carrying mostly technology-related content, seemed to be heading in the same direction.

An early exception to this paradigm was SunSITE-UNC, a joint project incorporated in 1992 and sponsored by Sun Microsystems and the University of North Carolina at Chapel Hill. The site's original administrators were quick to recognize the Internet's potential for information exchange, and served the first non-technological content on the World Wide Web: astronomy archives, and President Clinton's press releases. This was a relatively small start in what would become the site's public FTP and WWW archive, a sizeable collection of academic and not-so-acamedic content organized in a directory hierarchy by discipline: /pub/academic, /pub/books, /pub/multimedia, etc.

Rather than developing its collections like a physical library, ordering specific texts to fill its shelves, SunSITE administrators simply offered free web and FTP space to almost anyone who had something to say, and needed a venue in which to say it. The sole requirement was that the site adhere to the Collection Policy, which requires that the site further the teaching and research mission of the University, and be free of copyright or other

legal restrictions. For ten years the archive has cheerfully accepted the sites of the *International Union of Pure and Applied Chemistry,* Roger McGuinn's folk song archive, and Lou Sortam's *Big Candy Wrapper Ball* page alike.

Over the years SunSITE became MetaLab and MetaLab became ibiblio, but the inclusive, communal nature of the archive has remained unchanged – and its administrators' open attitude towards content management has created a unique situation among digital libraries. Although ibiblio employees retain administrative control of the machines, contributors are given as much publishing freedom as possible, with no space or content requirements, and with access to all of ibiblio's open-source infrastructure.

In ten years ibiblio has grown to have a significant impact on the University's network and on the greater World Wide Web: ibiblio traffic consistently accounts for nearly two thirds of UNC's outgoing bandwidth, and makes UNC one of few universities that generates more outgoing Internet traffic than incoming. Nearly 13,000 sites world wide link to ibiblio's homepage alone. Yet, no major bibliometric study of this archive has been undertaken. This study asks:

- How many collections does the archive hold?
- How old are the collections?
- What types or genres of sites are present?
- Which genres of sites take up the most space?
- How do these sets of numbers relate to one another?

More simply stated, what does ten years of free-range academic web publishing look like?

Expectations

ibiblio offers web space to its contributors with little or no obligation required of them, save three simple restrictions: no commercial activity, no porn, and no pirated music or software. Given the wide diversity of the archive's sites and its contributors, one might not expect common characteristics to emerge from the archive. However, if one were to venture a guess:

- **Genre:** ibiblio comprises around 800 proper collections, not including its sizeable Linux and Open Source archives. Given the site's medium and history, we could expect a high proportion of technology-related sites.
- **Age:** Given the site's inception, we may expect technology-related sites to be among the first. This may not be an indicator, however, of which genres of sites tend to be updated for the longest spans of time.
- **Size:** Sites with multimedia content should take up the largest physical disk space. We cannot presuppose which genre of site would contain the highest number of web pages.
- **"Freshness:"** Given the number of abandoned web sites on the commodity Internet, the online world's infamous short attention span, and the demanding workload common to many contributors with technical skills, we would expect a fair number of abandoned sites in the archive.

Limitations

In his *Evaluating Digital Libraries,* Marchionini tells us that "response time, storage capacity, transfer rate, user satisfaction, and cost per operation" might offer a technical

perspective on the physical performance of digital libraries, but would not be sufficient measurements in and of themselves (Marchionini 2000, p. 1). Unfortunately, not even all of this data is available for our examination.

There are two major sources of descriptive data that *are* available about the collections:

- the **Collection Index,** a contributor-maintained *nouveau* card catalog, and
- the archive's filesystem, which describes file dates, sizes, ownership, and permissions.

These two sources describe the sites from the "back end," where the site content's production takes place. The descriptive data we lack is on the "front end:" the access logs, which would track end-user access to the collections. It would be quite a feat for the archive to have retained all access logs, given the number of web sites and constant browsing of the archive over the past ten years. Unfortunately ibiblio only has access logs dating back to 1998, covering less than half of the life of the archive. For this reason, they are not included in this study.

Another major roadblock was the somewhat Utilitarian stance ibiblio administrators took during its formative years – the site was experimental, after all, and so the standards and procedures for adding and managing sites have changed significantly over time. In his *Evaluating Digital Libraries,* Marchionini tells us that "the effects of digital libraries will emerge over time as physical libraries, digital libraries and people mutually adapt and mature" (Marchionini 2000, p. 2). The results of this study might well represent the consequences of publishing in such a grass-roots, bottom-up manner.

During the site's formative years, administrators made great efforts to ensure that contributors were satisfied with the archive's resources in order for their sites to flourish and contribute to the archive's greater goals. After all, any strong society needs "citizens, not wage slaves" (Schuler 1996, p. 75). Despite noble efforts to ensure that each site is listed in the ibiblio Collection Index, many remain unlisted.

Several factors will confound our efforts to compare the sites equally (Turnbull 1996). Some web sites have been served through user home directories, and some web sites exist physically in one directory, yet may have pointers, or symbolic links, leading to them from other directories. Some sites are administered by several users who share group access on the machine, and some are controlled by one account which may be used by many people. All of these factors may affect the accuracy of the lists of dates, sizes, and other measurements we intend to take. Further, a number of ibiblio collections use databases for dynamically generated content.

Administrative access to the machines grants the freedom to collect as much data as possible. However, for the above reasons, certain data will simply be incomplete (such as the access logs which do not date back to the site's inception), some of it will be incorrect (such as errant time stamps due to filesystem corruption or software error), and some data was never kept in the first place (such as file creation time, which UNIX operating systems make no effort to keep). In the course of this examination it should be possible to identify and exclude *some* of these erroneous numbers. Unfortunately, many of them will remain undetectable – there are no other records against which we can compare.

Even if all of these problems could be rectified they would not be sufficient to describe a complete picture of the archive and its contributors. As Sylvan Katz warns us: “bibliometric indicators ... fall short [because] there is not a one-to-one match between publication output and research expenditure” and that “bibliometric indicators do not represent all publishing” (Katz 1999, p. 2).

Katz's points are well taken, barring access to complete and accurate data needed to perform usability, satisfaction, and other such performance reports on the archive. However, hopefully some semblance of contributor publishing patterns or other group behavior can be forensically constructed by examining the internal data which *is* available (Renninger 2002, p. 165).

II. Literature Review

A great deal of analysis has been conducted on many aspects of the World Wide Web – but bibliometric, social network, and community-oriented works are most closely related to this paper. Five selected works contributed significantly to this study and are discussed below.

Published in 1996, Don Turnbull's *Bibliometrics and the World Wide Web* was an early and prominent theory work on the emerging medium. Turnbull outlines how one might apply Information Science theory to the Web in order to measure documents in terms of citation, co-citation, and bibliographic coupling. This type of study can yield important results, including paying homage to pioneers, identifying related work, substantiating claims, and providing leads to poorly disseminated, poorly indexed, or un-cited work (Turnbull 1996).

However, Turnbull's article supposes that one merely has end-user access to these documents, and peers through its examination lens from this angle. It does not address the "back-end" bibliometric measurements which yield insight into the patterns and habits of web publishers. The back-end data collected for this study should yield some interesting results, considering the size and diversity of the ibiblio contributor base.

Bernardo Huberman's *Patterns in the Ecology of Information* (2001) follows Turnbull's lead in measuring links between documents in order to measure their connectedness.

However, Huberman goes on to apply high-level mathematical theories to measure traffic patterns and map out a topology of the Internet landscape. The text essentially superimposes Zipf's law of rarities and commonalities onto the "Ecology of Information" to show that very few Internet sites are very well connected, a small percentage of web sites are responsible for the largest percentage of actual traffic, and so on. Like Turnbull, Huberman relies on externally collected data for his study.

Gail Hodge's *Best Practices for Digital Archiving* (2000) laments the failure of modern web authors to preserve individual versions of works, including those which have been deleted, and is a call to arms for librarians and archivists to fend off the "new barbarians at the gate" in the name of preservation. Had ibiblio contributors more closely followed Hodge's recommendations over the years, our data would have yielded much more elaborate results for this study.

Jenny Preece's *Online Communities* (2000) discusses the establishment of online communities, and what Internet architects and new-world developers can do to facilitate communities on their sites. However, ibiblio is not a community in the sense of Preece's discussion – it is not a Yahoo Groups or other discussion site explicitly devoted to fostering individual communities, and many of its sites might co-exist for years without even being aware of one another.

In another sense ibiblio *does* function as a raucuous online community, powered by its contributors who may not work together or share common interests, but do share the sense that they are all making their voices heard by publishing their works in "the Public's Library" - which, as we will see, makes ibiblio more like a digital library than any other descriptor.

Doug Schuler's *New Community Networks – Wired for Change* is another early Internet work, published in 1996. In it Schuler offers his views on the current state of traditional communities in terms of economy, education, society, and other descriptive factors. He then outlines ways in which then-emerging technology could benefit mankind to form his idea of Utopia, often railing against legislators, psychologists, institutions, and manifestations of "the establishment." Because of his views of the possibilities Internet technology allows, I am including Schuler's work as a major literary reference – ibiblio exemplies several of these core values: offering a voice to those who might otherwise not be able to spread their message, encouraging and presenting opposing viewpoints with equal zeal, and offering its resources on request so that the site as a whole becomes a true collaborative effort.

III. Research Design and Procedures

ibiblio runs Red Hat Linux across all of its servers, so PERL scripts were native for parsing dates, ranking directories, and performing other data collection. The data comprised four descriptive sets:

- **UDC Category** – The ibiblio Collections Index classifies its sites in Universal Decimal Code (UDC) by primary and secondary categories. The Index's software package uses PHP and MySQL, so I was able to simply query the database machine and save the results to a local file for parsing. Collected data includes the collection ID, collection title, primary and secondary UDC categories for each collection, and whether the collection has been deleted (removed but not forgotten) or is marked historic (abandoned).
- **Timestamps** – Each file's timestamp is accessible via Linux's *stat* utility and will allow us to assess the archive's vitality when gathered. Unfortunately, file creation time was never kept and some timestamps may be wrong for several reasons – while we may be able to reject some as being obviously wrong, we have no choice but to accept the reported dates of the file's last modification at face value.
- **Directory age** – Using PERL scripts to aggregate and compare each file's timestamp, we can come up with the earliest file timestamp for directory (*minMod*), and the latest (*maxMod*). By comparing these two variables, we can discern the time span of each directory, and therefore the overall site. This might better be referred to as the directory's

"age," but this term would incorrectly imply using the present time as a reference point. This data is useless in cases of newer, dynamically generated sites, and confounded by sites which may be uploaded in one sitting and then not touched. However, it is my hope that my scripts can find enough of a range of ages to allow some insight into the publishing behavior of ibiblio contributors.

- **Directory size** – Again, we can use PERL and standard UNIX utilities to aggregate and sort the physical disk usage of directories (and therefore the greater collections).

All data was collected on January 10, 2003.

IV. Analysis of the Data

ibiblio is able to boast of being one of the oldest and largest academic repositories in the world, and the collected data does not disappoint us. The charts presented below reflect the measurements of 807 proper web collections taken from the site's Collection Index, and the dates and sizes of 2.34 million files and 111,000 actual directories. Note that there are directories and files beyond the domain of those comprising the proper collections – those of personal homepages, old or forgotten directories, etc. However, as the collections represent nearly all of the resources in the archives, the data below is really about them.

Timestamps

Of the numbers collected, the biggest surprise was probably the overall current file modification stamps, separated by year. Nearly half of the files were last updated during the year of 2002. Further, another 25% were modified in the years of 2001 and 2000, leaving the final quarter of timestamps trailing down to 1990. Although ibiblio was officially launched in 1992, some archives contain files dating farther back.

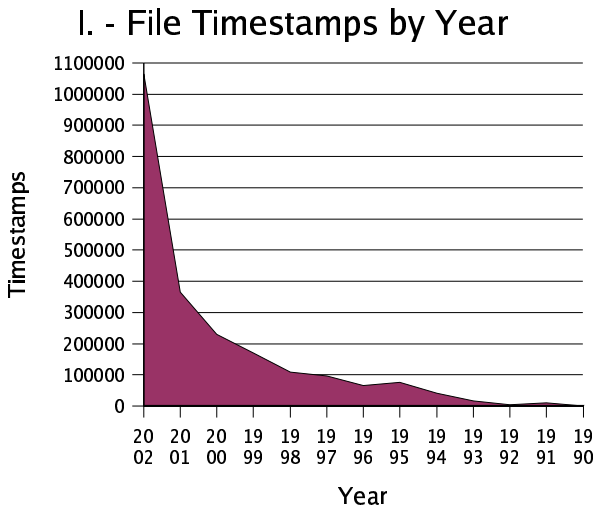

Illustration 1: File Timestamps by Year

Unfortunately the Collections Index records the creation date for very few collections and file creation times have been overwritten over the years, so we cannot calculate how the number of sites added per year affects the modification dates. In general, however, timestamps show a significant pattern towards “current” activity, so an appropriated Zipf's law could said to be in effect.

Directories

Likewise nearly 90% of the directory ages calculated from these timestamps were less than a year old, and 5% were between 1 and 2 years old. These directory numbers by themselves don't signify much, but when compared against the above timestamp results, they strongly suggest that, regardless of the rate of collection additions to the archive, its contents as a whole are constantly updated. This is especially surprising considering the minimal participation ibiblio requires of its contributors, and refutes the common notion that many web sites wind up abandoned or sparsely updated.

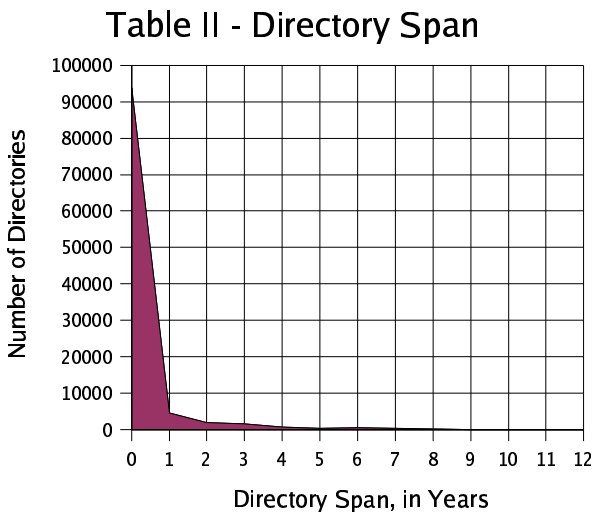

Illustration 2: Directory Span

UDC Categories

In keeping with the open philosophy of the site, ibiblio allows its contributors to classify their own collections. This hopefully bolsters the accuracy of the index, as no one

might best describe the collection as well as its maintainer. UDC collection categories on the site do not echo the wide discrepancy in percentages displayed by the timestamps, but do suggest that many more collections on the site are associated with the arts, history and social science than with religion and philosophy. Technology-related sites are ranked fourth by sheer number of collections, which is surprising given the site's early technology foundation and adherence to the open-source philosophy and standards. Please see Appendix A, Table IV for the actual list.

Again, these category measurements discuss the actual web collections, and do not include ibiblio's large Linux, Sun, and other software archives, which are a separate study unto themselves.

Directory Size by UDC

The ranked list of directory sizes becomes significant when the UDC category of each site is included in the list. The ranked list itself was around 110,000 lines, which I began ocularly parsing from the top, paring the list down to actual collections and removing redundant / unnecessary entries.

Although personal web sites don't *really* count as proper collections, the list of ranked directories would have been very inaccurate if they were simply ignored: personal sites account for more than a third of the physical disk space consumed in ibiblio's web space. As expected, the *Arts and Entertainment* collections consumed the most space as a category (much of this is legitimate MP3 content), followed by the nearly tied categories of *Technology and Applied Sciences* and *Geography, Biography, and History.* All other

sciences account for around 12% of the space, literature and reference account for another 7%, and religion, language and philosophy and psychology account for the remaining 3.5%. This makes perfect sense, as multimedia files guzzle disk space and text files do not. Please see Appendix A, Table V for the actual list.

UDC Categories vs. other descriptors

Given the ad-hoc, bottom-up publishing nature of the site, I hypothesized that it might be interesting to examine some of the above collection descriptors, if they were aggregated and examined according to the site's genre. I have compiled a listing of ibiblio's most popular collections, sorted by the collection's primary UDC heading, in groups of around ten per subject heading (see Appendix B). Unfortunately, no significant, discernible patterns emerge beyond the large multimedia presence. A co-worker pointed out to me that, since there is no central tendency among the collection descriptors, one can mathematically only treat the largest part of the "central" numbers as no more than noise (Efron 2003b). Conversely, this leaves the "outlier" collections for examination.

To compile this list, I compared the physical disk space consumed, the number of physical files in the collection, and the time span (not the age) of the collection. Outlier collections fell pretty easily into four categories: those which comprise only a small number of files and a small amount of disk space yet have a long life span, "small" collections which have a medium life span, collections which have a medium life span and are physically large, and collections which are both old and physically large.

Miniscule Collection, Medium Span:

UDC	Collection Title	Size	Files	Span
0	Libraries FAQ	0.49 MB	22	3 years
0	Newspaper Archs on the Web	1.28 MB	35	4 yrs, 9 mos.
500	Apiculture and Beekeeping	1.15 MB	44	4 yrs, 11 mos.

The above three collections display similar descriptive characteristics because they share the same function: all serve as portals to greater, more far-flung resources on the Internet. They each comprise a relatively small and equal number of files, very little disk space, and due to their function, the sites' spans are merely a measurement from their inceptions to the present time.

However, each site's organization and presentation reflects what one might expect from their respective maintainers. The Library FAQ site's resources are strictly categorized into an elaborate hierarchy, in what could almost be a miniature ontology, and the sparse graphics on the site are simple title banners and are common to all pages in the archive – very uniform and very organized, as would be expected of a library. The Newspaper Archives maintainers appear to be slightly more relaxed about its structure, presenting its links to archive resources worldwide in table format divided into continent – which properly reflects the geographical searching needs of the field, arguably the most useful for a Journalistic professional. The Apiculture and Beekeeping archive merely serves as a front end to the esoteric UseNet archive, with no organization and the content organized simply in chronological order – which conveys the notion that a technology-driven, tightly-organized website would be beyond the needs or the capability of those in such an organic profession.

Miniscule Collection, Broad Span

UDC	Collection Title	Size	Files	Span
100	alt.psychology.personality	0.14 MB	25	7 yrs, 2 mos.
100	Mind Uploading Home Page	0.29 MB	41	8 yrs, 2 mos.
500	Automated Weather Notification	0.16 MB	34	5 yrs, 7 mos.
600	Sustainable Farming	45 MB	1132	5 yrs, 7 mos.
700	EMUSIC-L/SYNTH-L	0.93 MB	928	8 yrs, 3 mos.
800	Internet Chinese Text	75 MB	3165	10 yrs, 7 mos.

Like the smaller sites discussed above, each of these sites are almost completely in text format, and most serve as portals to broader resources across the Net. However, the similarities end there.

Two such archives are stagnant – the EMUSIC-L list has been shut down and so the collection merely chronicles its 8 years of operation, and the personality archives can claim a span of 7 years only because the Keirsey foundation demanded that a third-party Keirsey temperament test be removed (and therefore artificially extending the directory's life span). Otherwise, the site's span would only be about two years, as it was abandoned when its maintainer graduated from college.

Joe Morris' Weather Site and the Chinese Text archive have each been well-maintained over the years and are good resources for their respective topics. Each offers a wide assortment of web references in addition to useful proprietary content.

The remaining archives can boast of their broad span for the right reasons – that the mailing list or newsgroup they're archiving is still active, or that their maintainers are diligently updating links and content. The Mind Uploading Homepage and Sustainable Farming site in particular are well-planned, elaborate archives which bely their small size.

Large Collection, Broad Span

UDC	Collection Title	Size	Files	Span
200	B-Greek	1,986 MB	208579	6 yrs, 2 mos.
200	Baha'i Page at MetaLab	659 MB	70651	7 yrs, 9 mos.
300	Nation of Gods and Earths	2621 MB	5850	8 yrs
400	Virtual Shtetl	393.6 MB	10716	8 yrs, 5 mos.
600	EcolandTech/InterGarden	633 MB	57512	6 yrs, 8 mos.
700	SITO	2420 MB	70021	10 yrs, 4 mos.
800	Project Gutenberg	16996 MB	18461	10 yrs, 10 mos.

Taken together, these sites represent some of the oldest and best-maintained collections on ibiblio. All but one of them offer nearly all of their physical disk space in proprietary content as opposed to offering links to other resources on the Web, are extremely well organized, and are wonderful academic resources on the Net.

However, two sites fall short of this Utopian model – the Nation of Gods and Earths and EcolandTech each offer extensive content, but have a somewhat relaxed organization that makes site browsing difficult and maintaining a sense of continuity while navigating the site impossible. Further, a good portion of the Nation of Gods and Earths' space is consumed by MP3s of questionable legality – which brings me to my final point: for all of these site's content and the obvious great effort put into developing and maintaining them, all but Project

Gutenberg could stand some Spring cleaning.

Large Collection, Medium Span

UDC	Collection Title	Size	Files	Span
900	DocSouth	3023 MB	34229	4 yrs, 4 mos.
600	Linux Documentation Project	584 MB	30965	6 yrs, 4 mos.
0	Special Libraries Assn News	135.2 MB	3835	5 yrs, 10 mos.

The DocSouth and Linux Documentation Projects and the SLA News sites echo the qualities of the better examples immediately above, but simply haven't been around for as long. Each are well-organized, easily navigable and offer a great deal of local content.

Two More:

Two of ibiblio's most popular sites not mentioned above, or even entered into the Collections Index, are Roger McGuinn's *Folkden* and Nic Pioch's *Webmuseum:*

UDC	Collection Title	Size	Files	Span
N/A	Folkden	223 MB	610	5 yrs, 3 mos.
N/A	WebMuseum	223 MB	4291	8 yrs, 7 mos.

The Folkden and WebMuseum are similar in that they have each been steadily and actively updated and augmented since their creation, and unlike most of the larger ibiblio sites listed above are primarily composed of multimedia content. As both sites have been conceived, created and maintained by one person, neither employ a strict organizational scheme or offer a consistent navigational interface. Regardless, to the ibiblio administrators' experience these are two of the most popular sites in the archive in terms of the exsting access logs and in user requests.

V. Discussion: On Virtual Communities, Social Networks and Digital Libraries

Put to its best use, "media *can* be useful. It need not be disconnected from society like a hallucinogenic dream world on the other side of the looking glass" (Schuler 1996, p. 220). He continues that while this noble service of giving voice to dischord may be a catalyst for culture clash, it should ultimately increase public knowledge and mutual understanding. In this sense, ibiblio could be argued to be a large virtual community in that it is composed of unsolicited, everyday contributors whose collections combine to form a Gestalt-esque system complete with the virtual community functions of discourse, accord and dischord.

However, the ibiblio community arguably might not measure up to Preece's definition of online community, which requires "people, purpose, policies, and computer systems" (Preece 2000, p. 10). Although each of these components are present and ibiblio contributors share the same purpose and computer systems, an internal citation analysis has shown that very few ibiblio sites refer to one another, and are held only to the minimal ibiblio Collection Policy – there is no sense of a proper virtual community, and there is no self-government (Efron 2003a, p.2).

The site hosts conflicting sites such as AtheistParents.org alongside the Carolina Crusade for Christ, and ibiblio's webmaster regularly receives complaints about slightly more controversial sites like the *Nation of Gods and Earths,* and the satirical *StayFree Magazine.* In this sense of collection dischord, ibiblio might be better described as a loose

social network populated by contributors who co-exist as “virtual strangers exchanging ideas and information” (Renninger 2002, p. 159). However, the effects of this separation boomerang in the form of user feedback, complaints, kudos, or proposed collaborations – echoing and continuing Sennett's views on a psychological aspect of public anonymity: “the lack of a strong, impersonal culture in the modern city ... has aroused a passion for fantasized intimate disclosure between people” (Sennett 1977, p. 255).

While one would readily admit that most of the communication between contributors and users could hardly be considered intimate, the sheer amount of user mail sent to the webmaster and other role accounts (estimated at 600 per month) indicates that ibiblio end-users world-wide take every opportunity to applaud, chide, or question things they find on the ibiblio site. To appropriate the saying, “the more things change, the more they stay the same,” the more diverse and varied the collections become, the more traffic, communication, and activity the archive attracts – building community in the sense of shared space and shared purpose. As one researcher states: “traditional content reappears in a sort of pastiche wherein new media refer to old, and vice versa. Rather than a radical break, then, there is a continuous loop of influence” (Shields 1996, p. 127).

John Dewey once argued that the ideas of fraternity, liberty, and equality were hopeless abstractions if they were separated from community existence, and as described above, ibiblio contributors could easily be argued to tacitly contribute to such a community (Dewey 1927). In this sense ibiblio's contributors' leveraging the medium of the Web in order to disseminate their largely non-technical, “real world” content reinforces their citizenry in this new community. Rather than substantiating technophobic fears that the mainstreaming of

the Internet would segregate and separate society even further, “[it] serendipitously brings to us, in our living rooms and offices, a sense of connectedness, but it is an aimless connectedness, a kind which reassures that between 'us' and 'them' there may be a some common ground after all” (Jones 1997, p. 17).

For all its community functions, at heart ibiblio is simply a vast and diverse digital library, rightfully powered by its contributing members. One well-respected expert on digital libraries has stated that “evaluating digital libraries is a bit like judging how successful is a marriage” (Marchionini 2001, p. 1). In this case, a three-way marriage between the site's technology infrastructure, contributors and end users.

From a technical perspective, the ibiblio project has certainly proven itself over the years. Once running on an infrastructure of only one server and now about to move to a distributed web cluster of nearly 20 machines, the archive accounts for enough network traffic that the University's network monitoring tools now ignore it: the ibiblio cluster consistently generates so much traffic that it confounds troubleshooting the rest of the network via the tools' visual graphs and pie charts.

For all its growth and success, the archive does little in the scope of collections development. This is where the contributors take over: administrators rarely recruit other Internet sites and most new collections on the archive are referred by web searches or word-of-mouth. Unlike those in many other digital libraries, ibiblio contributors maintain both their collections and their collections' metadata in the site's Index.

While the archive may not be accessible to the extreme of Richard Brautigan's fictional library (1966), to which anyone could add whatever collection they wanted, about whatever they wanted, wherever they wanted, for the above reasons ibiblio may boast of being highly accessible to anyone with a computer and an area of interest – truly a Public's Library.

VI. Summary and Implications

Initial measurements of dates, sizes and genres of the proper collections and additional directories revealed some surprising results – namely, that one of the Internet's oldest and largest public archives suffers little of the stagnation and "collection rot" that would be expected of such an organization.

Further research on this topic should compare the above "back-end" measurements against the end-user usage indicated by the site's access logs. One would expect, for example, that a few collections would be responsible for most of the site's traffic, echoing the clustering patterns outlined in Adamic's *Small World Web* (1999). Unfortunately, ibiblio only retains access logs back to 1998, and the sheer size of the log files is prohibitive to an unaccomplished programmer's examination.

Another research strategy might be to measure "in-links" to construct an internal citation analysis, which might provide insight into ibiblio site clustering. Measuring the number and concentrations of broken links might also provide some insight into individual collections' integration with one another and with the Internet.

At ten years old, ibiblio remains one of the oldest and largest academic repositories on the Internet. As a relatively unstudied entity, it holds a great deal of information about the academic web publishing of today as well as ten years ago – and awaits inquiry.

References

Adamic, L. (1999). The Small World Web. Springer, New York.

Brautigan, Richard. (1966). The Abortion: An Historical Romance. Pocket Books, New York.

Dewey, John. (1927). The Public and its Problems. Henry Holt, New York.

Efron, Miles. (2003a). Link Attachment (Preferential or Otherwise) in Contributor-Run Digital Libraries. Submitted to the Joint Conference on Digital Libraries, May 2003. Houston, TX.

Efron, Miles. (2003b). Personal correspondence: mefron@ibiblio.org.

Hodge, Gail. (2000). Best Practices for Digital Archiving. Retrieved June 19, 2002, from http://www.dlib.org/dlib/january00/01hodge.html.

Huberman, Bernardo. (2001). The Laws of the Web: Patterns in the Ecology of Information. MIT Press, Cambridge, MA.

ibiblio.org Collection Policy. (2003) Retrieved on January 3, 2003 from http://www.ibiblio.org/collection.html.

Jones, Steven, Ed. (1997). Virtual Culture – Identity and Communication in Cybersociety. SAGE Publications, London.

Katz, J. Sylvan. (1999). Bibliometric Indicators and the Social Sciences. Retrieved February 3, 2003 from http://www.sussex.ac.uk/Users/sylvank/pubs/ESRC.pdf.

Marchionini, Gary. (2000). Evaluating Digital Libraries: A Longitudinal and Multifaceted View. Retrieved December 12, 2002, from http://www.ils.unc.edu/~march/perseus/lib-trends-final.pdf.

Preece, Jenny. (2000). Online Communities: Designing Usability, Supporting Sociability. John Wiley & Sons, Ltd. Chichester.

Renninger, K. Ann and Shumar, Wesley, Eds. (2002). Building Virtual Communities – Learning and Changing in Cyberspace. Cambridge University Press, Cambridge, UK.

Schuler, Doug. (1996). New Community Networks – Wired for Change. Addison-Wesley, New York, New York.

Sennett, Richard. (1977). The Fall of Public Man. Alfred A. Knopf, New York.

Shields, Rob, Ed. (1996). Cultures of the Internet: Virtual Spaces, Real Histories, Living Bodies. Sage Publications, London.

Turnbull, Don. (1996). Bibliometrics and the World-Wide Web. Retrieved December 11, 2002, from http://donturn.fis.utoronto.ca/research/bibweb.html.

Bibliography

Egghe, Leo and Rousseau, Ronald. (1990). Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science. Elsevier Science Publishers, Amsterdam.

Hafner, Arthur. (1989). Descriptive Statistical Techniques for Librarians. American Library Association, Chicago and London.

Hoadley, Irene and Clark, Alice, Eds. (1972). Quantitative Methods in Librarianship: Standards, Research, Management. Greenwood Press, Westport, CT.

Kubicek, Herbert, Dutton, William, and Williams, Robin, Eds. (1997). The Social Shaping of Information Superhighways – European and American Roads to the Information Society. St. Martin's Press, New York.

Nicholas, David and Ritchie, Maureen. (1978). Literature and Bibliometrics. Clive Bingley LTD, London.

O'Neil, Dara. (2001). Merging Theory with Practice: Toward an Evaluation Framework for Community Informatics. Retrieved January 15, 2003 from http://maven.gtri.gatech.edu/~doneil/aoir.pdf.

Rheingold, Howard. (1993). The Virtual Community: Homesteading on the Electronic Frontier. MIT Press, Cambridge, MA.

Segaller, Stephen. (1998). Nerds 2.0.1 – A Brief History of the Internet. TV Books, New York, New York.

Smith, Marc and Kollock, Eds. (1999). Communities in Cyberspace. Routledge Press, London.

Appendix A: Granular bibliometric Data

Table II: Current File Modification Stamps by year (as of January 10, 2003)

year	number	percent
2003	40,968	1.75%
2002	1,065,317	45.50%
2001	365,947	15.25%
2000	230,587	9.85%
1999	171,265	7.31%
1998	108,826	4.65%
1997	97,216	4.15%
1996	65,005	2.78%
1995	76,290	3.26%
1994	41,670	1.78%
1993	16,246	0.69%
1992	5,281	0.23%
1991	10,449	0.47%
1990	1,338	0.06%
	2,296,405	97.73%
from	2,341,356	<-- discrepancies due to parsing problems.

Table III: Directory Span via Filestamps Breakdown

years	numdirs	%
12	20	0.02%
11	0	0.00%
10	35	0.03%
9	50	0.05%
8	153	0.15%
7	413	0.39%
6	484	0.46%
5	398	0.38%
4	762	0.73%
3	1636	1.56%
2	2028	1.94%
1	4661	4.45%
0	93828	89.62%

255 dirs older than ten years due to comp-sci archive which was likely tarred from another machine.

Table IV: ibiblio Collections by UDC Primary Category
(not counting personal sites or software archive)

Rank	UDC	Primary Category	Count	Percent
1	700	Arts and Recreation	144	18.49%
2	900	Geography, Biography and Hist.	142	18.23%
3	300	Social Sciences	137	17.59%
4	600	Technology, Applied Sciences	118	15.15%
5	500	Natural Science and Math	70	8.99%
6	000	Reference	56	7.19%
7	800	Literature	38	4.88%
8	200	Religion and Theology	29	3.72%
9	400	Language	23	2.95%
10	100	Philosophy and Psychology	22	2.82%
			779	100.1%

Table V: Top 400 Largest Collections by UDC Category

udc	num	percent	
PRS	137	34.25%	<-- personal sites
700	77	19.25%	<-- arts and entertainment
600	48	12.00%	<-- technology, applied sciences
900	44	11.00%	<-- geography, biography and history
300	29	7.25%	<-- social science, education
500	21	5.25%	<-- natural science and math
800	17	4.25%	<-- literature
000	13	3.25%	<-- reference
200	8	2.00%	<-- religion and theology
400	6	1.50%	<-- language
100	0	0.00%	<-- philosophy and psychology
	400	100%	

Appendix B: Selected Collection / UDC Data

UDC	Collection Title	Size	Files	Span
0	Bartlett Collection	28.136 MB	1020	1 year
0	Catalog of Copyright Entries	3.53 MB	100	1 month
0	CCCC Bibliography	10.53 MB	118	4 yrs, 2 mos.
0	Extimacy	64.93 MB	927	10 mos.
0	Libraries FAQ	0.49 MB	22	3 years
0	NC Electronic Public Info	0.916 MB	96	7 mos.
0	Newspaper Archs on the Web	1.28 MB	35	4 yrs, 9 mos.
0	Special Libraries Assn News	135.2 MB	3835	5 yrs, 10 mos.
0	UNC JOMC Library	529 MB	3985	7 yrs, 7 mos.
0	World Population Counter	73.89 MB	3187	8 yrs, 4 mos.

UDC	Collection Title	Size	Files	Span
100	alt.psychology.personality	0.14 MB	25	7 yrs, 2 mos.
100	Context	0.71 MB	56	4 yrs, 2 mos.
100	Dreamerforum	4.95 MB	763	1 yr, 4 mos.
100	InnerPeace - Free Self-Help	11.58 MB	1100	1 yr, 6 mos.
100	Lunar Institute of Technology	72.90 MB	3187	8 yrs, 5 mos.
100	Mind Uploading Home Page	0.29 MB	41	8 yrs, 2 mos.
100	Lenhart's Philosophy Archive	0.85 MB	12	7 mos.

UDC	Collection Title	Size	Files	Span
200	Atheist Parents	39.56 MB	4177	2 yrs
200	B-Greek	1,986 MB	208579	6 yrs, 2 mos.
200	B-Hebrew	0.20 MB	2	3 yrs, 1 mo.
200	Baha'i Page at MetaLab	659 MB	70651	7 yrs, 9 mos.
200	Buddhism - Metaxu	4.66 MB	484	7 mos.
200	Christian Phenomenology	0.22 MB	10	1 yr, 8 mos.
200	Corpus Paulinum	0.96 MB	14	1 yr, 4 mos.
200	Gospel of Mark Discussion	0.20 MB	9	1 yr, 5 mos.
200	Sockpa Kangtsen at Gaden	5.93 MB	146	3 yrs
200	Zen @ MetaLab	0.71 MB	99	4 yrs

UDC	Collection Title	Size	Files	Span
300	Actualidad Colombiana	10.71 MB	669	2 yrs, 11 mos.
300	Charity Legal Aid	0.48 MB	10	1 mo.
300	Creative Commons	162.17 MB	2784	1 yr
300	Cuba	9 MB	207	7 yrs, 2 mos.
300	NC Million Mom March	9.62 MB	91	1 yr, 5 mos.
300	Nation of Gods and Earths	2621 MB	5850	8 yrs
300	Online Burma Library	783 MB	2564	1 yr, 4 mos.
300	Socialist Party Cybercenter	12.77 MB	120	5 yrs, 5 mos.
300	Tibetan Multi-Education	2.26 MB	58	2 mos.
300	Widen the Web	2.87 MB	189	1 yr, 2 mos.

UDC	Collection Title	Size	Files	Span
400	Annali d'italianistica	6.53 MB	90	1 mo.
400	Fowler Anglo-Saxon Lexicon	28.14 MB	1020	11 mos.
400	Gaelic Homepage	9.34 MB	298	7 yrs, 8 mos.
400	Internet Chinese Text Archive	75.39 MB	3165	10 yrs, 6 mos.
400	Little Greek 101	5.36 MB	311	2 yrs, 10 mos.
400	Open Translation Engine	4.03 MB	145	7 mos.
400	Virtual Shtetl	393.6 MB	10716	8 yrs, 5 mos.
400	Vocabumonkey in Dutch	1.06 MB	253	1 mo.

UDC	Collection Title	Size	Files	Span
500	Apiculture and Beekeeping	1.15 MB	44	4 yrs, 11 mos.
500	Astrobiology	10.96 MB	646	1 yr, 7 mos.
500	Automated Weather Not.	0.16 MB	34	5 yrs, 7 mos.
500	BOTNET	567.89 MB	5595	4 yrs, 8 mos.
500	IUPAC	1017 MB	13013	3 yrs, 8 mos.
500	Lineberger NACF	326.79 MB	16386	5 yrs, 3 mos.
500	NC Int. Dark-Sky Assn	34 MB	150	2 yrs, 11 mos.
500	Permaculture	12.47 MB	268	10 yrs
500	Plant Information Center	377 MB	3254	2 yrs
500	UNC Herbarium	19.1 MB	408	3 yrs, 9 mos.

UDC	Collection Title	Size	Files	Span
600	AIFIA	41 MB	2755	2 yrs, 3 mos.
600	Cafe au Lait	95 MB	6608	7 yrs, 10 mos.
600	EcolandTech/InterGarden	633 MB	57512	6 yrs, 8 mos.
600	FreeDOS Project	583 MB	4034	2 yrs, 10 mos.
600	Henriette's Herbal Homepage	535 MB	23751	5 yrs, 6 mos.
600	Linux Documentation Project	584 MB	30965	6 yrs, 4 mos.
600	Permaculture Activist	0.19 MB	38	3 yrs, 1 mo.
600	Sustainable Farming	45 MB	1132	5 yrs, 7 mos.
600	Thinking in C++ and Java	4.14 MB	21	1 yr, 2 mos.
600	VINCENTweb	21 MB	992	4 yrs, 6 mos.
600	Virtual Cell	330 MB	4174	3 yrs, 2 mos.

UDC	Collection Title	Size	Files	Span
700	Ball	3 MB	140	2 yrs, 7 mos.
700	Bawdy Ballads and Folklore	27 MB	254	4 yrs, 7 mos.
700	EMUSIC-L/SYNTH-L	0.93 MB	928	8 yrs, 3 mos.
700	Elvis Presley Home Page	17 MB	817	7 yrs, 11 mos.
700	Fela Project	2.4 MB	180	1 yr, 3 mos.
700	Flicker Films	906 MB	433	7 mos.
700	Philm Freax	16 MB	667	6 yrs, 7 mos.
700	Propoganda [sic]	999 MB	13291	3 yrs, 10 mos.
700	SITO	2420 MB	70021	10 yrs, 4 mos.
700	'Sup	400 MB	682	3 yrs, 4 mos.
700	UNC Student Television	1455 MB	2686	4 yrs, 6 mos.
700	Vietnam Multimedia Archive	4.35 MB	422	5 yrs, 11 mos.

UDC	Collection Title	Size	Files	Span
800	Arthur Miller Society	1.55 MB	48	3 yrs, 7 mos.
800	Dr. Fun Archive	661 MB	11937	9 yrs, 1 mo.
800	Internet Chinese Text	75 MB	3165	10 yrs, 7 mos.
800	Internet Poetry Archive	859 MB	751	4 yrs, 5 mos.
800	La Palabra	69 MB	29	1 mo.
800	NC Native Am. Storytelling	144 MB	47	3 yrs, 7 mos.
800	Open Book Project	675 MB	8549	1 yr, 7 mos.
800	Paul Green	250 MB	128	1 yr, 10 mos.
800	Project Gutenberg	16996 MB	18461	10 yrs, 10 mos.
800	Walker Percy Project	149 MB	23992	7 yrs, 6 mos.

UDC	Collection Title	Size	Files	Span
900	DocSouth	3023 MB	34229	4 yrs, 4 mos.
900	Duke Homestead	177 MB	325	6 yrs, 5 mos.
900	Hollerin' Contest	91 MB	87	3 yrs, 5 mos.
900	Hyperwar	894 MB	15961	6 yrs, 6 mos.
900	LA Slaves	29 MB	544	10 mos.
900	Pearl Harbor	145 MB	2694	5 yrs, 10 mos.
900	Skysoldier	130 MB	5774	4 yrs, 2 mos.
900	Southern Oral History Project	100 MB	282	3 mos.
900	SouthNow	58 MB	614	1 yr, 10 mos.
900	VolunteerTibet	20 MB	1233	2 yrs, 7 mos.