Working Paper Series
ISSN 1170-487X

Applications for Bibliometric Research in the Emerging Digital Library

by Sally Jo Cunningham & Mahendra Vallabh

Working Paper 95/17
May 1995

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Applications for Bibliometric Research in the Emerging Digital Library

Sally Jo Cunningham, Mahendra Vallabh
Department of Computer Science
University of Waikato
Hamilton, New Zealand
email: sallyjo@waikato.ac.nz, mike@waikato.ac.nz

Abstract: A large amount of research literature has recently become available on the Internet through "digital libraries". This migration of information from paper to electronic media promises to have a huge impact on the way that research is performed, as documents become more widely, cheaply, and quickly distributed than is possible through traditional publishing. A secondary use for these document repositories and indexes is as a platform for bibliometric research. We examine the extent to which the new digital libraries support conventional bibliometric analysis, and discuss shortcomings in their current forms. Interestingly, these electronic text archives also provide opportunities for new types of studies: generally the full text of documents are available for analysis, giving a finer grain of insight than abstract-only online databases; these repositories often contain technical reports or pre-prints, the "gray literature" that has been previously unavailable for analysis; and document "usage" can be measured directly by recording user accesses, rather than studied indirectly through document references.

1. Introduction

In recent years a number of "digital libraries" have become available through the Internet. These libraries are generally repositories of one or two types of document (technical reports, journal articles, pre-prints, or conference proceedings), grouped by discipline. The full text of documents are often available, as well as bibliographic records. They are maintained by professional societies, universities, research laboratories, and even private individuals. Access is free, both to search and to download documents.

The emergence of these subject-specific digital libraries is particularly important given the pattern of access to materials presently employed by research scientists. Informal exchanges of preprints, reprints, and photocopies of papers passed on by colleagues currently are major venues for the transmission of scientific information between researchers. In a study of how scientists locate and retrieve documents, the dependence on these sources ranges from 12% (for chemistry) to 39% (for mathematics) of all papers cited in researchers' own publications [Hallmark, 1994]. For specialized subjects such as high energy physics, this dependence on informal or extra-library dissemination can be much higher. Ginsparg [1994a-b] reports that fields in physics have traditionally relied heavily on preprint exchanges, and the digital repositories of preprints begun in 1991 have to a large extent supplanted conventional publishing and physical paper mailing of technical reports. By providing ready access to information sources that are already preferentially utilized by scientists, the digital libraries show potential to increase access to information that until recently was expensive or difficult to acquire in paper form. Indeed, in some fields (most notably physics) this process has already begun, as researchers in less developed countries report access to ongoing

research through the Internet repositories that their local libraries could not afford to acquire through conventional journal subscriptions [Ginsparg, 1994a-b].

The primary use for new bibliographic resources is, of course, for the contents of the documents involved. A secondary use for emerging resources is as a basis for bibliometric analysis of the subject field. As the name implies, *bibliometrics* is the quantitative analysis of the bibliographic features of a body of literature; it is used to gain insight into how scholars in a field disseminate information and to detect historical trends in research methods and topics. Typically, bibliometric methods involve accumulating statistics on the structural attributes of documents (shelf length, number of words, number of illustrations, etc.), general characteristics of a body of literature (obsolescence rate, development of research fronts, "important" documents and researchers, etc.), and citation analysis (tracing relationships between documents through their referencing of each other).

Bibliometric studies were pioneered early this century. The first work, published in 1917, measured the relative contributions of different countries to the fields of comparative anatomy between the years 1543 and 1860 [Cole, 1917]. However, the sheer difficulty of accumulating statistics discouraged bibliometric research until the advent of large bibliographic databases in the 1960's. These resources include the online databases held by vendors such as Dialog and Prodigy; citation databases such as the Science Citation Index and Social Science Citation Index developed by the Institute for Scientific Information; and various subject-specific databases, both public and private. These computerized bibliographic databases sparked a significant increase in the number of large-scale bibliographic studies, as significant portions of the collection and analysis of data could be automated [Hawkins, 1977; McGhee, 1987]. The availability of CD-ROM versions of bibliographic databases has been of particular importance, since they provide a cheaper alternative to the online commercial databases [Burton, 1988].

These computerized bibliographic resources have drawbacks, however. The greatest is that the full text of documents are rarely available, and even abstracts are not always present. This obviously limits the types of bibliometric research that can be conducted *solely* through these databases. In addition, these databases are generally limited to formally published documents (those appearing in selected books, journals, and conference proceedings). The "gray literature" of technical reports, pre-prints, and other works not formally published are largely ignored, and it is this absence of easy access to these documents that has hampered the analysis of these important forms of scientific communication.

The digital libraries currently in existence complement the online and CD-ROM bibliographic databases by addressing these shortcomings. The digital libraries are best suited for examinations of the "physical" characteristics of documents (document length, number of documents, number of authors, etc.), usage studies (geographic or institutional origin of users, date/time of access, individual patterns of document retrieval, etc.), and sociology of science research (origin and development of subject disciplines, diffusion of research, etc.). However, co-citation and bibliographic coupling research is not well-supported, and conducting these studies requires considerable effort on the part of the researcher.

The variety of bibliographic repositories in the available digital libraries in itself has great potential in conducting bibliometric research. Sigogneau [1991] presents a case study illustrating the ways in which the strengths of different databases can be played off each other; they conduct a fine-grained analysis of the emergence of research fronts in molecular and cellular biology, and demonstrate that the observations gleaned from two complementary bibliographic databases provide greater insight into their problem. Similarly, it appears that the types of bibliographic data that can be gleaned from the

relatively unstructured digital libraries can be profitably combined with data from online databases, CD-ROMS, and other more conventional bibliographic resources.

This paper is organized as follows: Section 2 discusses the types of indexing and searching available with current digital libraries; Section 3 gives examples of conventional bibliometric techniques applied to Internet-accessible archives; Section 4 discusses opportunities to directly measure usage of documents; and Section 5 presents our conclusions.

2. Overview of digital libraries

A number of Internet-accessible searching and indexing systems currently exist. We will concentrate on representative systems that provide access to scientific literature, including: the physics E-PRINT ARCHIVE, which has supplanted journals or pre-print mailings as the primary information dissemination point for several areas of physics [Ginsparg, 1994a-b]; the Unified Computer Science Technical Report Index (UCSTRI) in Indiana; the NTRS system, which provides an index for NASA technical reports [Nelson, 1994]; the WATERS distributed database of computer science technical reports [Maly, 1994]; the DIENST repository, indexer, and search engine that currently provides an interface to a handful of computer science technical report repositories [Davis, 1994a-c]; Carnegie Mellon's MERCURY search engine (again, currently applied to computing literature); and the HARVEST tools for information gathering, index construction, and searching [Bowman, 1994a-b]; and a computer science technical report server based on the *mg* full text indexing software [Witten, 1994, 1995]. A list containing these and other subject- or site-specific research report servers is maintained at [NASA].

types of indexing and search fields

At present, the types of searching available on most systems are limited. Most schemes index on user-supplied document descriptions, abstracts, or similar document surrogates. UCSTRI, for example, provides a searchable text index based on text obtained by parsing the index file that is present by convention in most ftp directories of technical reports. This text does not necessarily characterize the report very closely, and in any case is only a small subset of the full text in the document. Moreover, the parsing procedure is sensitive to the format of the index file, and cannot be guaranteed to succeed. HARVEST's ESSENCE sub-system [Hardy, 1993; Hardy, 1994] extracts "content summaries" whose composition may vary widely. ESSENCE relies on filetype-specific procedures to extract relevant information from the document itself; for example, LaTeX documents can be parsed for author and title information. ESSENCE's success in extracting an appropriate document surrogate depends on its ability to cope with the file type of the document and the semantic cues provided by that type. The *mg*-based system is unique in that it indexes the full text of documents by extracting the text from common file formats (LaTeX, PostScript, etc.). The remaining systems under consideration – DIENST, NTRS, WATERS, and the physics E-PRINT ARCHIVE – require the submitter of the technical report to provide cataloging information, while WATERS requires a designated site librarian to maintain a local catalog.

The search interfaces for existing digital libraries are more primitive than those ordinarily found in online bibliographic databases or library catalogs. UCSTRI, HARVEST, and the proposed *mg*-based system primarily provide keyword searches, as their indices do not contain formal bibliographic catalogs. The DIENST, NTRS, WATERS, and physics E-PRINT ARCHIVES can support more detailed information about each report (such as author, title, and subject searching), but this more sophisticated search functionality comes at the expense of requiring participating repositories to use specific software. As a consequence, these latter systems provide access to only a handful of sites, whereas UCSTRI, HARVEST, and the *mg*-based system can access a broad range of providers.

effect of indexing on bibliometric research

Existing digital libraries have a variety of shortcomings for bibliometric applications:

- lack of consistency in field formatting.

  Current digital libraries usually acquire bibliographic information from either the authors of submitted articles or automatic extraction routines. Neither of these methods produce records with standard formatting, which causes problems with automated bibliometric analysis. Consider the following examples selected from entries in the hep-th (high energy physics) archives:

  (i) Authors: A. Yu. Alekseev, V. Schomerus
  (ii) Authors: Adel Bilal and Ian. I. Kogan
  (iii) Authors: Paul S. Aspinwall and David R. Morrison (with an appendix by Mark Gross)
  (iv) Authors: A. H. Chamseddine and Herbi Dreiner (ETH-Zurich)

  In this case, typical for existing digital libraries, there is no standardized format for authors' names (here, appearing with full names, initials plus last name, and a mixture of the two); no standard convention for separating author names (here, either a comma or "and" are used); and parenthetical information can include a variety of information such as the name of an associate author or the institutional affiliations of an author. Manual processing or specially crafted software would be required to reformat these fields for analysis.

- duplicate entries.

  Digital libraries that "harvest" items from a variety of sources may contain duplicate items. The irregular formatting of the bibliographic information makes it difficult to automatically detect these duplicates.

- implicit field tagging

  In some repositories, items are not explicitly tagged with certain types of information – most commonly the document's date of publication or production. Instead, the date is implicit in the document's title (eg, its numeration in a technical report series) or in the location of the document in the file structure of the repository (eg, separate directories exist for each year). A second common piece of implicit data is the authors' institutional affiliations. This may be contained in the document itself (typically on a cover page), or may be implicit in the document's location (for example, a corporation's technical reports are stored in its ftp repository). Again, in these cases special processing is required to append this field information to a document record for bibliometric analysis.

Given these limitations, the following steps are required to construct a file useable for automated bibliometric analysis [modified from Burton, 1988]:

- construct a search strategy that will isolate an appropriate subset of the digital library

- retrieve the selected citations and, if necessary, the associated document files

- add missing fields as necessary (date of publication/production, country of origin, etc.), and convert fields to a standard format

- extract text from document files. Some digital libraries provide ascii files. Otherwise, utilities are available to convert common file formats such as TeX, LaTeX, and PostScript to ascii.

- identify and delete duplicates. Depending on the source repository chosen, duplicates may occur within a single digital library as well as when search results are combined from several archives.

- If necessary, parse document text files to extract information for bibliometric analysis. Some formats, such as LaTeX, can be readily parsed to identify author, title, etc. For other formats and other types of information, such as locating the references of documents, ad hoc methods are required.

It is likely that many of these problems will be addressed as the Internet-based document indexing systems mature. Even minor changes can greatly increase the useability of a bibliographic database for bibliometric research. For example, the addition of an explicit date tag to many online databases in 1975 sparked new applications in time series research [Burton, 1988].

3. Opportunities for applications of bibliometric techniques

One type of bibliometric research concentrates on quantifying fundamental, structural details about a subject literature: how many items are published, how many authors are publishing, over what time period documents are likely to be used, etc. More complex studies analyze the relationships between documents: how documents cluster into subjects, how "invisible colleges" are formed, etc. The following examples give a flavour of the bibliometric research that is possible using the emerging digital libraries:

- examining the "physical" characteristics of archived documents.

  One simple type of bibliometric study characterizes the formats of different literatures. For example, Figure 1 presents a the range of the size of computer science technical reports as measured by their word count. 188 technical reports were sampled from Internet-accessible repositories[1]. Note that the number of words falls into an approximately normal distribution, and that the range in number of words is quite high (between 1,000 and 38,000 words).

  This type of analysis is of particular interest for technical reports, since they have not been studied in the same detail as formally published papers. A comparison of the physical characteristics of the formal and informal literature could provide supporting evidence for common beliefs about the relationship between the two types of documents (eg, do publishing constraints force journal and proceedings articles to be shorter than technical reports, and therefore presumably omit technical details of findings? do technical reports contain more/less extensive reference sections? if the reference sections are smaller, does that indicate that scientists tend to "research first", and do literature surveys later?).

---

[1]Documents were randomly sampled from the DEC (ftp://crl.dec.com/pub/DEC/CRL/tech-reports/), Sony (ftp://ftp.csl.sony.co.jp/CSL/CSL-Papers), and Ohio (ftp://archive.cis.ohio-state.edu/pub/tech-report/) technical report repositories

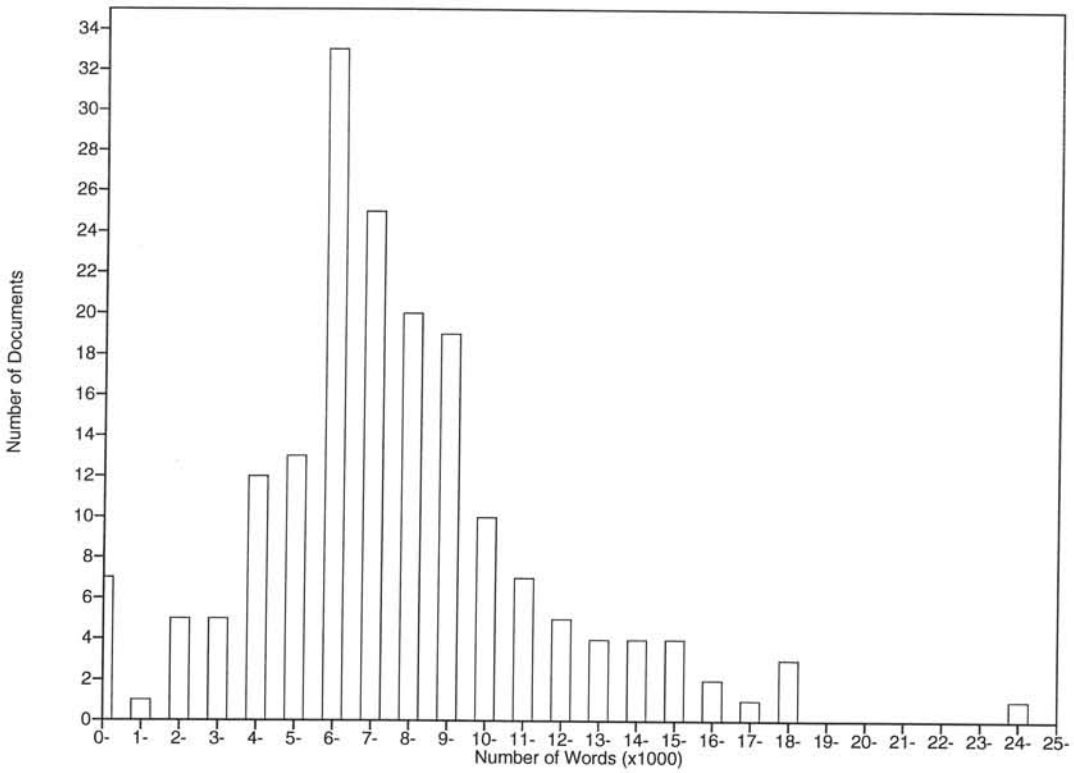

Figure 1. Range of sizes of CS technical reports, measured by number of words

- obsolescence studies.

A document is considered obsolete when it is no longer referenced by the current literature. Typically, documents receive their greatest number and frequency of citations immediately after publication, and the frequency of citation falls rapidly as time passes. One technique for estimating the obsolescence rate of a body of literature– the *synchronous* method – is to find the median date in the references of the documents. This median date is subtracted from the year of publication for the documents, yielding the *median citation age*. As would be expected, this median varies between the disciplines. Typically the social sciences and arts have a higher median citation age than the "hard" sciences and engineering, indicating that documents obsolesce more quickly for the latter fields.

As noted in Section 2, references are not generally explicitly tagged in existing digital repositories. However, reference dates can usually be extracted from the document text by first locating the reference section (usually delimited by a "references" or "bibliography" section heading), and then extracting all numbers in the appropriate ranges for dates for the field under study.

To illustrate this process, the 188 technical reports described in the section above were used as source documents for a synchronous obsolescence study. Conveniently, these repositories organize technical reports into sub-directories by their date of publication. The reference dates for each technical report were extracted by scanning the document's file for numbers of the form 19XX, since previous studies indicate that few if any computing reports reference documents published in previous centuries [Cunningham, 1995]. Table 1 presents the median citation age calculated for these documents, broken down by repository and the year of publication for the source documents from which the reference dates were extracted:

repository	year of source document publication	median citation age (years)
DEC	1990	2
	1991	3
	1992	2
	1993	3
	1994	3
Sony	1992	2
	1993	3
	1994	4
Ohio	1992	3
	1993	3
	1994	3

Figure 2. Median citation ages for technical report repositories

The median citation age ranges between 2 and 4 years, which is consistent with previous examinations of computing and information systems literature. Graphs of the distribution of reference dates exhibit typical obsolescence patterns (see, for example, a graph of the Ohio reference dates in the Appendix). The graph of each repository shows the exponential curve found in obsolescence studies, including the final droop due to an "immediacy effect" as fewer very new documents are available for citation [Price, 1970].

- time series/trend analysis

The rate of growth for a body of literature can be estimated by measuring the accumulation of items in an associated bibliographic database [Hall, 1989]. A simple technique is to plot the number of items produced in each year of interest. Figure 3, for example, presents a distribution of references in the computer science bibliography maintained by Alf-Christian Achilles[2]. It shows a typical exponential increase in the number of items produced as time passes, with the "droop" at the most recent year most likely an artefact of a time lag in adding documents to the collection.

A similar analysis of a carefully selected portion of a bibliographic database can provide a detailed picture of a specific research area [McGhee et al, 1987]. Typically, a keyword or fielded search is used to break a broad-range database into component subfields, and then the publication trends in these sub-fields are analyzed. Since many digital libraries do not tag author, institution, etc. data, some types of analysis are difficult to automate.

In addition, most digital libraries do not currently extend far enough into the past to allow meaningful analysis over an extended time period. However, the *rate* of accumulation can be estimated as the number of items added in a short period such as a year.

---

[2]The bibliography is located at:
URL: ftp://ftp.cs.umanitoba.ca/pub/bibliographies/index.html
The distribution of references depicted in Figure 3 is located at:
ftp://ftp.cs.umanitoba.ca/pub/bibliographies/Statistics.html

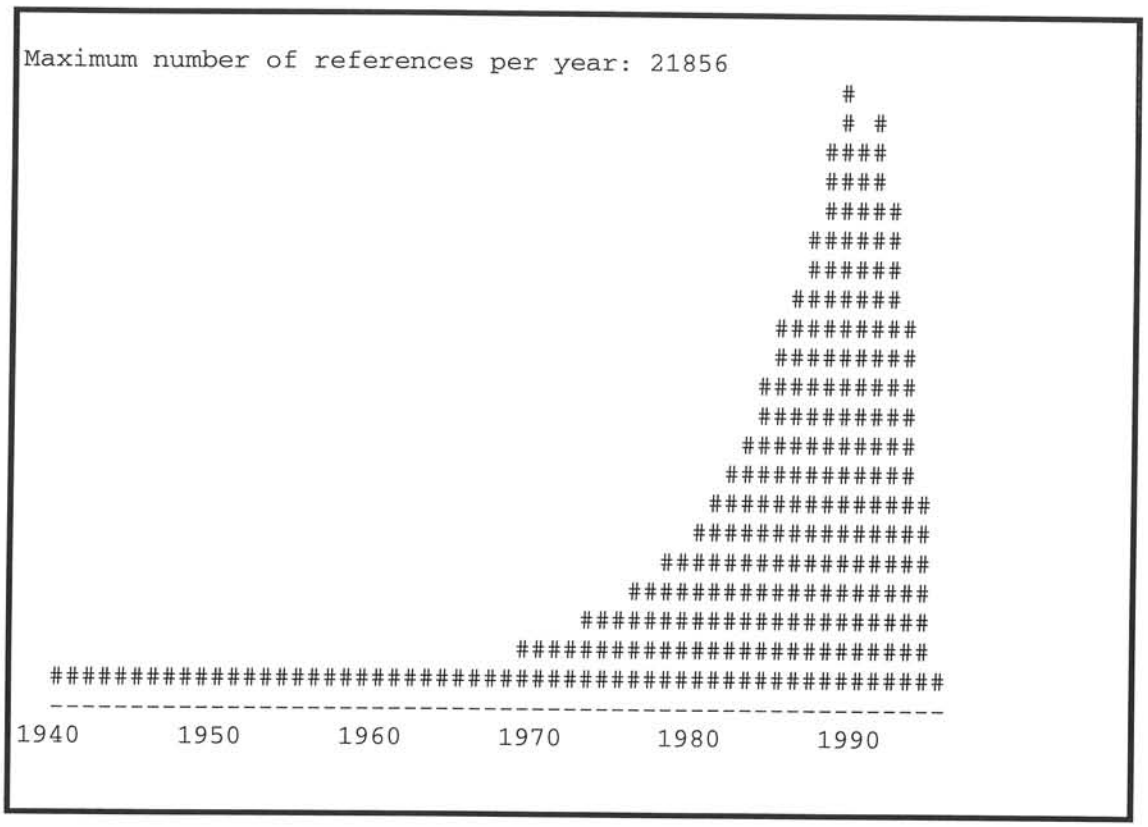

Figure 3. Distribution of the items in a computer science bibliography over time

- cocitation and bibliographic coupling studies

  The rate at which documents cite each other (cocitation) or cite the same documents (bibliographic coupling) can be used to produce "maps" of a subject literature. These techniques rely on analysis of the references of documents, and these references must be in a common format. While digital libraries contain full text of documents, their references are not standardized, and indeed are not even tagged as such. To perform these studies the references must be manually extracted and processed–a tedious process that is only worthwhile for documents (such as technical reports) that are not included in existing citation databases such as the Science Citation Index and Social Science Citation Index.

- detecting cycles or regularities in the rate of production of research

  Analysis of trends in the production of technical reports can give indications about working conditions that affect research; for example, is more research produced over the summer, when the teaching load is lighter? or is research steadily produced throughout the year?

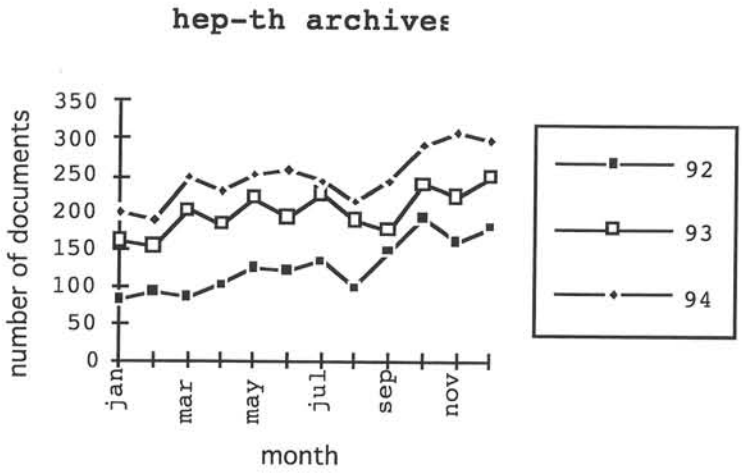

Figure 4. Distribution of the number of documents submitted to hep-th, 1992-1994

Figures 4 and 5 present statistics on document accumulation in the hep-th (high energy physics) e-print server. This system is one of the oldest formal pre-print archives, and has become the primary means for information dissemination in its field. Examination of these figures reveals several trends. Clearly the absolute number of documents deposited in the repository has tended to increase over the time period. For all three years, research production has its lowest point in January and February, increases through May and June, then decreases until August and September. At that point the rate of production steps up, reaching a yearly peak in November and December. This pattern is less clear for 1992, which might be expected as the archive was established in mid-1991.

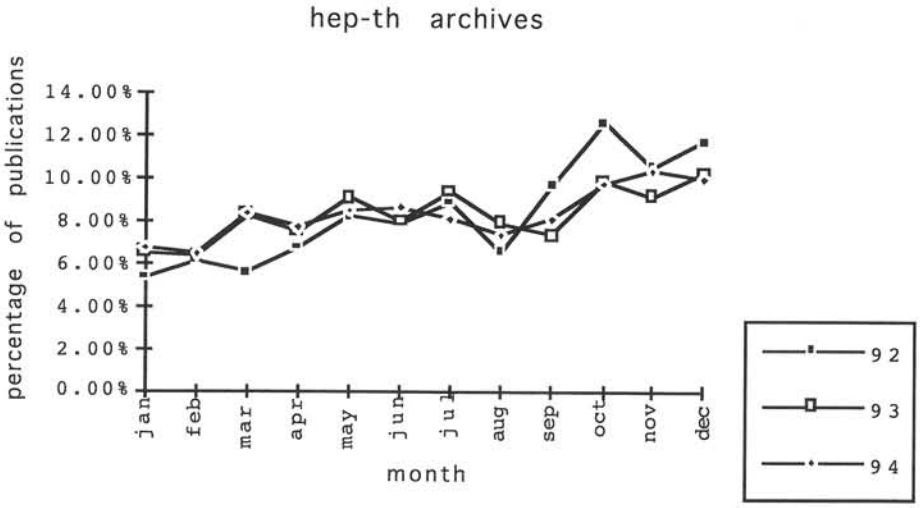

Figure 5. Distribution of the percentage of documents submitted to hep-th, 1992-1994

4. Analysis of usage data

The emerging Internet-based digital libraries will permit research on scientific information collection and use at a much finer grain than is possible with current paper libraries or online bibliographic databases. Current bibliometric or scientometric research of this type must measure information use indirectly – for example, through examination of the list of references appended to published articles. By monitoring accesses to digital libraries, document usage can be measured directly, and a more detailed picture can be painted of how scientists pursue their research.

Surprisingly few previous reports on index or repository projects provide even a cursory analysis of the usage data they collect on their systems. As a sample of the types of analysis possible, Paul Ginsparg notes a seven day periodicity in the number of search requests made to the physics e-print archives (Figure 6, reproduced from Ginsparg 1994a). From this he adduces that many physicists do not yet have weekend access to the Internet (an alternative, slightly more cynical hypothesis is that even high energy theoretical physicists take the weekend off).

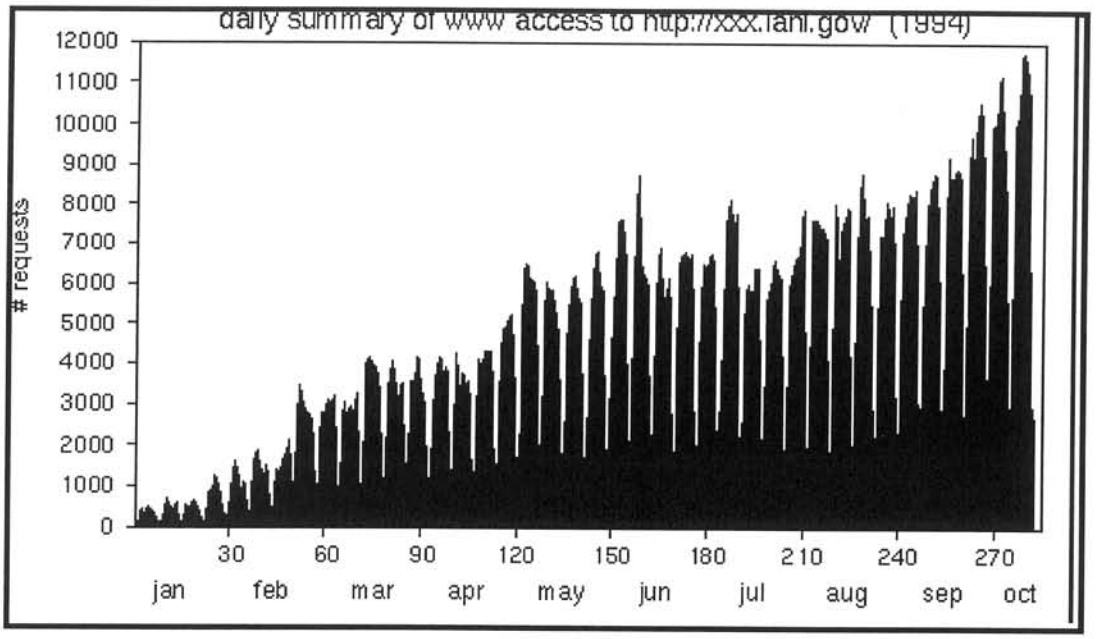

Figure 6. Summary of search requests to the physics pre-print archives

Analysis of search requests by geographic location, institution, and even individual user are also possible. Figure 7 presents a portion of the summary of usage statistics (sorted by domain name) for the computer science bibliography collection maintained by Alf-Christian Achilles[3]. Summaries of this type could be invaluable for studies of geographic diffusion and distribution of research topics.

The *mg*-based computer science technical report indexing system will attempt a detailed examination of the research activities of its target users: the New Zealand academic computing researchers. The size of this research community is particularly appropriate for an intensive study, as it is limited to seven universities. In addition to monitoring user access times, this system will also record the specific documents retrieved by users. Analysis of the index terms for these documents will allow the creation of user and site profiles that characterize the types of research carried on in these departments.

[3]URL: ftp://ftp.cs.umanitoba.ca/pub/bibliographies/index.html

```
HTTP Server Domain Statistics

Covers: 07/01/94 to 10/07/94 (99 days). All dates are in local time.
Total unique hosts: 26527
Total number of requests: 642328
1 level, sorted by domain name, 69 unique domains.

# reqs : Last Access (M/D/Y) : Domain
---------------------------------------------

 36734 : 10/07/94 : (numerical domains)
   259 : 10/07/94 : Argentina (.ar)
   672 : 10/07/94 : Old style Arpanet (.arpa)
  2475 : 10/07/94 : Austria (.at)
  7218 : 10/07/94 : Australia (.au)
  3463 : 10/07/94 : Belgium (.be)
  1996 : 10/07/94 : Brazil (.br)
 23866 : 10/07/94 : Canada (.ca)
 27101 : 10/07/94 : Switzerland (.ch)
   230 : 10/06/94 : Chile (.cl)
  1670 : 10/07/94 : China (.cn)
    96 : 10/07/94 : Colombia (.co)
 27049 : 10/07/94 : US Commercial (.com)
   281 : 10/07/94 : Costa Rica (.cr)
   217 : 10/07/94 : Czech Republic (.cz)
 46310 : 10/07/94 : Germany (.de)
```

Figure 7. Usage statistics for a computer science bibliography

5. Conclusion

This study suggests opportunities for conducting bibliometric research on the evolving digital libraries. These repositories are suitable platforms for conventional bibliometric techniques (such as obsolescence studies, quantification of physical characteristics of documents comprising a subject literature, time analysis, etc.). The ability to directly monitor access to documents in digital libraries also enables researchers to explicitly quantify document usage, as well as to implicitly measure usage through citations. Additional facilities could aid in the performance of bibliographic experiments, such as: improved tagging of document fields; provision of utilities to strip out titles, authors, etc. from common document formats; and the ability to easily eliminate duplicate entries from downloaded library subsets. Unfortunately, the most useful of these additional facilities – those associated with a higher degree of cataloging – run counter to the underlying philosophy of many digital libraries: to avoid all manual processing of documents. While adherence to this principle means that users cannot explicitly search by author, date, keyword, etc., it permits the constructors of the digital libraries to provide access to as many documents as possible, without restricting the library contents by requiring documents to be in any special format.

The digital libraries complement the information currently available through paper, online, and CD-ROM bibliographic resources. While these latter databases generally have the advantage of standardized formatting of bibliographic fields, the digital libraries are freely accessible, often contain "gray literature" that is otherwise unavailable for analysis, and generally make the full text of documents available. The insights gained from analysis of digital libraries will add to the store of "information about information" that we have gained from older types of bibliographic repositories.

References

Blythe, J: "On-line CS Tech reports".
<URL:http://www.cs.cmu.edu:8001/afs/cs.cmu.edu/user/jblythe/Mosaic/cs-reports.html>

Bowman, C., Danzig, P., Hardy, D., Manber, U., & Schwartz, M., 1994a: Harvest: A scalable, customizable discovery and access system, *Technical Report CU-CS-732-94*, Department of Computer Science, University of Colorado, Boulder, Colorado.
<URL:ftp://ftp.cs.colorado.edu/pub/cs/techreports /schwartz/Harvest.ps.Z>

Bowman, C.M., Danzig, P.B., Manber, U., and Schwartz, M.F., 1994b: Scalable Internet resource discovery: Research problems and approaches, *Communications of the ACM 37(8)*, pp. 98-107.

Burton, Hilary D., 1988: Use of a virtual information system for bibliometric analysis, *Informaton Processing & Management 24(1)*, pp. 39-44.

Cole, P.J., and Eales, N.B., 1917: The history of comparative anatomy, *Science Progress 11*, pp. 578-596.

Cunningham, S.J., and Bocock, D., 1995: Obsolescence of computing literature, to appear in *Scientometrics*. Also available as *Working Paper Series 95/8*, Department of Computer Science, University of Waikato (Hamilton, New Zealand).

Davis, J. & Lagoze, C., 1994a: Dienst, a protocol for a distributed digital document library, Internet Draft (work in progress).
<URL:http://cs-tr.cs.cornell.edu/Info/dienst_protocol.html>

Davis, J. and Lagoze, C., 1994b: "Drop-in" publishing with the World Wide Web, *Proceedings of the Second International WWW Conference*, Chicago.
<URL:http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Pub/davis/davis-lagoze.html>

Davis, J. & Lagoze, C., 1994c: A protocol and server for a distributed digital technical report library, *Technical Report 94-1418*, Computer Science Department, Cornell University.
<URL:http://cs-tr.cs.cornell.edu/TR/CORNELLCS:TR94-1418>

Ginsparg, P., 1994a: After dinner remarks: 14 Oct '94 APS meeting at LANL.
<URL: http://xxx.lanl.gov/blurb>

Ginsparg, P., 1994b: First steps towards electronic research communication, *Computers in Physics 8(4)*, p. 390-401.

Hallmark, J., 1994: Scientists' access and retrieval of references cited in their recent journal articles, *College and Research Libraries 55(3)*, pp. 199-210.

Hardy, D., and Schwartz, M.F., 1993: Essence: A resource discovery system based on semantic file indexing, *Proceedings of the USENIX Winter Conference*, p. 361-374.

Hardy, D., Schwartz, M., 1994: Customized Information Extraction as a Basis for Resource Discovery, *Technical Report CU-CS-707-94*, Department of Computer Science, University of Colorado, Boulder, Colorado. To appear in

an upcoming issue of *ACM Transactions on Computer Systems*. <URL:ftp://ftp.cs.colorado.edu/pub/techreports /schwartz/Essence.Jour.ps.Z>

Harris, Rik: "Computer Science Technical Reports Archive Sites." <URL:http://www.rdt.monash.edu.au/tr/siteslist.html>.

Hawkins, D.T., 1977: Unconventional uses of on-line information retrieval systems: on-line bibliometric studies, *Journal of the American Society for Information Science 28*, pp. 13-18.

Maly, K., Fox, E.A., French, J.C., and Selman, A.L., 1994: Wide area technical report server, *Technical Report* , Dept. of Computer Science, Old Dominion University. <URL: http://www.cs.odu.edu/WATERS/WATERS-paper.ps>

McGhee, P.E., Skinner, P.R., Roberto, K., Ridenour, N.J., and Larson, S.M., 1987: Using online databases to study current research trends: an online bibliometric study, *Library and Information Science Research 9*, pp. 285-291.

NASA: "Technical Reports, Preprints and Abstracts." (list of sites supporting digital libraries or indexing services) <URL: http://www.larc.nasa.gov/org/library/abs-tr.html

Nelson, M.L., Gottlich, G.L., and Bianco, D.J., 1994: World Wide Web implementation of the Langley Technical Report Server, *NASA Technical Memorandum 109162*, Langley Research Center, Hampton, Virginia. <URL: ftp://techreports.larc.nasa.gov/pub/techreports/larc/94/tm109162.ps.Z>

Price, D.J. de Solla, 1970: Citation measures of hard science, soft science, technology, and nonscience, Nelson, C.E., and Pollock, D.K., ed., *Communication among scientists and engineers*, Heath Lexington.

Salton, G., and McGill, M.J., 1983: *Introduction to modern information retrieval*, McGraw-Hill Book Company.

Sigogneau, M.J., Bain, S., Courtial, J.P., and Feillet, H., 1991: Scientific innovation in bibliographical databases: a comparative study of the Science Citation Index and the Pascal database, *Scientometrics 22(1)*, pp. 65-82.

VanHeyningen, M., 1994: The Unified Computer Science Technical Report Index: Lessons in indexing diverse resources, *Proceedings of the Second International WWW Conference*, Chicago. <URL: http://www.cs.indiana.edu/ucstri/paper/paper.html#ref-odlyzko>

Witten, I., Moffat, A., and Bell, T., 1994: *Managing Gigabytes: Compressing and indexing documents and images*, van Nostrand.

Witten, I.H., Cunningham, S.J., Vallabh, M., and Bell, T.C., 1995: A New Zealand digital library for computer science research, *Working Paper 95/6*, Department of Computer Science, University of Waikato, Hamilton (New Zealand).

Appendix: Distribution of ages of referenced for Ohio technical reports