# Applications for Bibliometric Research in the Emerging Digital Libraries

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

**Abstract:** Large numbers of research documents have recently become available on the Internet through "digital libraries", and these collections are seeing high levels of use by their related research communities. 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 "grey 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. While the technology promises in the future to support large, heterogenous collections, at present the most widely used of the academically-focussed digital libraries are generally repositories of one or two types of document (typically technical reports, journal articles, pre-prints, or conference proceedings), grouped by discipline.

A distinguishing characteristic of these digital libraries is that the full text of documents are often available for retrieval, as well as bibliographic records.The sciences are represented much more heavily in the present crop of digital libraries than the social sciences, arts, or humanities. They are maintained by professional societies, universities, research laboratories, and even private individuals. Access is generally 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 the sciences. In one study, the dependence on these sources ranges from 12% (for chemistry) to 39% (for mathematics) of all papers cited in researchers' own publications [11]. A qualitative study of study of how computer scientists locate and retrieve documents (computing is one of the domains considered later in this paper) indicates that for that field, technical reports and research documents found in various locations on the Internet are a preferred source of information [6]. Many of the digital library systems discussed in this paper are repositories for just this type of literature. The documents tend to be of high quality: primarily technical reports or working papers from research institutions (both academic and commercial), as well as advance copies of work accepted for publication in conventional paper journals. Moreover, these digital libraries are also coming to include refereed work published digitally (in electronic journals). Anecdotal evidence suggests that in their fields, these digital libraries are coming to be the resource of choice for locating cutting edge work.

For specialized subjects such as high energy physics, this dependence on informal or extra-library dissemination can be much higher. Ginsparg ([9], [10]) reports that fields in physics have traditionally relied heavily on preprint exchanges, and the digital repositories of physics preprints begun in 1991 (the PHYSICS E-PRINT ARCHIVES) 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 ([9], [10]).

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. With the conventionally published scientific literature, the sheer difficulty of accumulating statistics discouraged bibliometric research until the advent of large bibliographic databases in the 1960's. 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 ([12], [13]). 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 [3].

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 "grey 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. They are best suited for examinations of the "physical" characteristics of documents (for example, document length), analysis based on

bibliographic information that can be automatically extracted from the document text or the sometimes unevenly formatted bibliographic records (such as obsolescence studies), and usage studies (geographic or institutional origin of users, date/time of access, individual patterns of document retrieval, etc.). Because references are present in the document file but not identified by field, 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 et al [15] present 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 to detect information-seeking patterns in researchers; and Section 5 presents our conclusions.

## 2. Indexing and searching in current digital libraries

At present, the types of indexing fields for most academically-oriented digital library systems are limited. Many schemes index on user-supplied document descriptions, abstracts, or similar document surrogates (for example, the PHYSICS E-PRINT ARCHIVE [10], a collection of physics pre-prints and technical reports). As will

be discussed below, the quality of this user-provided data can be highly variable, and may unfavorably impact the usefulness of the index for searching. Alternatively, a designated site librarian may maintain a catalog (eg, the WATERS [14] system, now subsumed by NCSTRL (`http://www.ncstrl.org/`), both primarily collections of computer science technical reports); in this case the quality of the bibliographic information may be expedited to be higher, but fewer sites will be likely to support such a librarian and therefore fewer documents are likely to be included in the digital library. In a "harvesting" system such as the computer science technical report collections supported by HARVEST [2] or the NEW ZEALAND DIGITAL LIBRARY computer science technical report collection ([16], [17]), documents are indexed from passive repositories (that may not even be aware that their documents are being included in the digital library). Harvesting systems therefore cannot rely on the presence of bibliographic data of any sort.

Because of the relative paucity of high-quality bibliographic data available to many of the current academically- or research-focussed digital library collections, their search interfaces tend to be more primitive than those ordinarily found in online bibliographic databases or library catalogs. Systems such as NCSTRL can support 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 may provide access to a small number of sites than harvesting systems. Harvesters may access a broader range of providers, but at the penalty of being limited to unfielded, keyword searches over the raw text of the document or document surrogate.

Specifically, the indexing in existing digital libraries has a variety of shortcomings for bibliometric applications:

- *lack of fielded indexing:* As noted above, some large and widely used digital libraries (such as the computer science technical report collection of the NEW ZEALAND DIGITAL LIBRARY) may lack formal cataloging entirely, and rely on

keyword searching over the raw document text. Obviously this makes field-dependent analysis more difficult (for example, locating documents produced by specific authors), and in the worst case my require a manual examination of all files in the collection in order to reliably identify a desired document subset. However, keyword search techniques that approximate fielded searching results may suffice: for example in the NEW ZEALAND DIGITAL LIBRARY computer science technical report collection, limiting the keyword search for "Johnson" to a search of first pages only is likely to retrieve documents written by Johnson (since for the majority of computer science technical reports, the first page contains little more than author, title, date, and institution details).

A more principled approach to extracting bibliographic information is embodied in the CiteSeer tool [1]. This software parses raw, unfielded academic documents and attempts to identify such indexing information as author, title, reference list, etc. Obviously such a tool cannot attain 100% accuracy over a heterogenous document collection, but in practice it appears useful in that it can make a good first pass in processing a set of documents, providing an initial set of parsed documents for analysis. The remaining (presumably much smaller) set of unparsable documents can then be dealt with manually.

- *lack of consistency in field formatting:* Current digital libraries usually acquire bibliographic information from either the authors of submitted articles or automatic extraction routines (retrieving bibliographic details from catalog files that may or may not be in a given document site, and that may or may not be in an easily parsable form). 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) collection of the PHYSICS E-PRINT 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 draw documents from a variety of sources may inadvertently contain duplicate items. Unfortunately, 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.

- *extraction of document text:* Few of the documents stored in the research-oriented digital libraries discussed in this paper are straight ascii text; instead, documents may appear in a variety of file formats, such as LaTeX, PostScript, PDF, etc. If the contents of the documents are to be automatically processed (for example, to count the words in a document, or to extract reference publication dates for an obsolescence study), then the text must be extracted. Utilities are available to convert most common document formats to ascii.

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 [3].

## 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, such as how documents cluster into subjects. 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 relatively straightforward 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 length in pages. Of the 45,720 documents in the CSTR collection as of April 1998, nearly 1600 did not contain page divisions in their files (and hence are excluded from analysis). Note that the number of pages in the shorter documents (<50 pages) falls into an approximately normal distribution (slightly skewed to the left), while presumably the longer documents represent Masters' and Doctoral theses. A surprising number of documents are very short (between one and 5 pages); these may represent the type of condensed results frequently found in the "technical notes", "short papers", and "poster sessions" of computing conferences and journals. The average number of pages per document, 27.5, appears to be slightly longer than the common upper bound for a computing journal article, although this observation must be confirmed by a similar study of the lengths of formally published computing articles.

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. For example, 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 reference sections of technical reports are longer than those of published articles, then citation links are being ommitted in published works; if technical reports contain fewer references, then this may confirm earlier indications that computer scientists tend to "research first" and do literature surveys later [6].

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

*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, 188 technical reports were sampled from Internet-accessible repositories[1] and used as source documents for a synchronous obsolescence study. Conveniently, the repositories chosen organize technical reports into sub-directories by their date of publication. The reference dates for each technical report were automatically extracted by software that scanned 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 [5]. 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:

Table 1. 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 ([5], [4]). When graphed, the distribution of reference dates show the exponential curve typically found in obsolescence studies, including the final droop due to an "immediacy effect"

as fewer very new documents are available for citation [7]. These types of results provide confirmation that references used in computer science technical reports (the pre-eminent “grey literature” of the computing field) conforms to the same patterns as references found in the formally published literature.

*co-citation and bibliographic coupling studies*

The rate at which documents cite each other (co-citation) 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 2. Distribution of the number of documents submitted to hep-th, 1992-1994

Figures 2 and 3 present statistics on document accumulation in the hep-th (high energy physics) e-print server, a part of the PHYSICS E-PRINT ARCHIVE. 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 3. 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. However, it is well known that authors do not necessarily include in the reference list all documents that could have been cited, and conversely that not all references listed may have been actually "used" in performing the research; citation behavior can be affected by a number of motivating factors (Garfield lists *15* possible reasons in [8]).

Digital library transaction logs provide a powerful tool for direct analysis of document "usage": since digital libraries contain the actual document (rather than only a document surrogate), the relative amount of "use" that a digital library's clients make of a given document sees can be estimated from the number of times the document file is downloaded (and, presumably, the document is read). Note that file downloading is a much stronger statement on the part of the user than, for example, having a bibliographic record appear in the query result set for a conventional bibliographic system; the user downloads only *after* the document has been found potentially relevant through examination of its document surrogate. Additionally, downloading is frequently time-consuming and sometimes costly (depending on local pricing for

Internet access). Downloaded documents are therefore highly likely at least to be scanned, if not read closely. The transaction logs for a digital library can provide a global picture of the use of documents in the collection, since all user interactions with the library can be automatically logged for analysis. By contrast, it is of course impossible to track usage of print bibliographies, and very difficult to monitor usage of bibliographic data available on CD-ROM across more than one or two sites.

Furthermore, analysis of search requests by geographic location, institution, and sometimes even individual user are also possible. As an example, Table 2 presents a portion of the summary of usage statistics (broken down by domain code) for queries to the computer science technical collection of the NEW ZEALAND DIGITAL LIBRARY. Examination of the data indicates that the heaviest use of the collection comes from North America, Europe (particularly Germany and Finland), as well as the local New Zealand community and nearby Australia. As expected for such a collection, a large proportion of users are from educational (.edu) institutions; surprisingly, however, a similar number of queries come from commercial (.com) organizations, indicating perhaps that the documents are seeing use in commercial research and development units.

Table 2. Accesses to the NEW ZEALAND DIGITAL LIBRARY CS collection by Domain Code

Of course, usage levels can also be further broken down by IP number (indicating institutions), and systems requiring users to register may also be able to analyze usage on an individual basis. Since the query strings themselves are also recorded in the transaction logs, this domain/institution/individual activity could also be linked to specific subjects through the query terms. Summaries of this type could be invaluable for studies of geographic diffusion and distribution of research topics.

Transaction log analysis can also indicate time-related patterns in the information seeking behavior of digital library users. As a sample of this type of analysis, Paul Ginsparg notes a seven day periodicity in the number of search requests

made to the PHYSICS E-PRINT archives (Figure 4, reproduced from [9]). 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 4. Summary of search requests to the physics pre-print archives

## 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, if possible, manual processing and formal cataloging of documents. While adherence to this principle can limit the accuracy of fielded searching (or indeed, preclude it altogether), it can also avoid the cataloging bottleneck and permit digital libraries to provide access to larger numbers of documents.

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 "grey 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


[1] Bollacker, K.D., S. Lawrence, and C.L.Giles, CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications, *Proceedings of the Second International Conference on Autonomous Agents* (Minneapolis/St. Paul, May 9-13), 1998.

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

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

[4] Cunningham, S.J., An empirical investigation of the obsolescence rate for information systems literature, *Library and Information Science Research*., 1996, http://library.fgcu.edu/iclc/lisrissu.htm

[5] Cunningham, S.J., and D. Bocock, Obsolescence of computing literature. *Scientometrics 34(2)* (1995), pp. 255-262.

[6] Cunningham, S.J. and Lynn Silipigni Connaway, Information searching preferences and practices of computer science researchers, *Proceedings of OZCHI '96* (1996) 294-299.

[7] de Solla Price, D.J., Citation measures of hard science, soft science, technology, and nonscience. In: C.E. Nelson and D.K. Pollock (eds), *Communication among scientists and engineers* (Heath Lexington, 1970).

[8] Garfield, E., *Citation Indexing: Its theory and application in Science, Technology and Humanities (*Wiley, 1979).

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

[10] Ginsparg, P., First steps towards electronic research communication, *Computers in Physics 8(4)* (1994) 390-401.

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

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

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

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

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

[16] Witten, I.H., S.J. Cunningham, M. Vallabh, and T.C. Bell, A New Zealand digital library for computer science research, *Proceedings of Digital Libraries '95* (1995) 25-30.

[17] Witten, I.H., C. Nevill-Manning, and S.J. Cunningham, A public library based on full-text retrieval, *Communications of the ACM* 41(4), 1998, p. 71

[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