

Library Statistics: reflecting yesterday, today and tomorrow

Cathie Jilovsky
CAVAL Collaborative Solutions

ABSTRACT: *A consistent and reliable yet evolving data collection is an essential tool for modern library management. Statistical data can be used to support management decisions, in benchmarking and in making submissions for grant monies. Data however is being needed more and more for purposes that are different to what was originally envisaged. The focus has moved from purely recording past activity levels to the provision of data for library assessment, service quality and business intelligence methodologies.*

This paper will examine the issues in maintaining a consistent and reliable data collection that remains relevant and usable in the changing information environment. Approaches to achieving a balance between the necessity to record past performance and to ensure data compatibility over time whilst incorporating new and evolving measures for today and into the future will be explored. The future will involve cross-sectoral uses of data and also measures for information delivery which we are yet to imagine, yet alone confront. The author's experience with the collection, management and publication of the Australian and New Zealand Academic Library Statistics (CAUL Statistics) will provide some practical examples of the issues discussed.

Libraries have collected statistics for many years. Many aspects of the data collection and reporting processes have changed over time, including the details of and reasons for gathering particular data elements, the methods employed for collecting the data, and the way in which the statistics are used. The focus has moved from purely recording past activity levels to the provision of data for library assessment, service quality and business intelligence methodologies.

It is important that the data collection is consistent and reliable so that it can be used with confidence. Statistical data may be used in many different ways, such as supporting management decisions, for benchmarking with other libraries and in making submissions for grant monies. Data however is being needed more and more for purposes that are different to what was originally envisaged.

This paper will outline and discuss issues relevant to maintaining library statistical data, with particular reference to the annual collection and publication of the CAUL Statistics.

Why collect statistics?

Historically libraries have counted those things that are easy to count such as the number of items in the collection, the number of loans and the number of books catalogued. Although these types of data do describe and record aspects of the library and its operations, they are not necessarily helpful with planning issues like future space requirements and staffing needs, and may be almost useless for assisting with decisions relating to implementing new technologies.

Statistics may simply record the size and activity of a library at a point in time, but they can also provide the data for benchmarking, for planning and demonstrating value. Time-series data, i.e. data collected for the same library (or group of libraries) using the same data elements over a number of years, is often used to illustrate change or stability.

Many libraries include statistical data in public documents such as newsletters and annual reports, this is a way of illustrating the magnitude of the library service to customers and/or stakeholders. However, as will be discussed in detail at this conference, nowadays this is considered only part of the process of library assessment. Other measurement tools, particularly feedback from users on the services they expect and how they perceive these are being delivered, are commonly used to collect additional data. Qualitative data (i.e. descriptive data containing comments collected using methods such as surveys and focus groups) adds extra dimension and depth to quantitative data (i.e. statistics). The use of multiple measurement tools will facilitate validation and hence increased confidence in the results.

Academic Libraries are increasingly being measured as part of wider University evaluations, such as the RAE (Research Assessment Exercise) in the United Kingdom. In Australia, the RQF (Research Quality Framework) is designed to measure the quality and impact of all publicly funded research carried out in Australia's Universities and in New Zealand, the PBRF (Performance Based Research Fund) has been designed to encourage & reward research excellence (this was evaluated in 2003 and is scheduled again for 2006).

What data should be collected?

Items that are easy to count or those that have been counted in the past may not be what is needed to answer the questions that need to be answered today. So it is vital that the statistical elements are regularly reviewed and updated in order to ensure that relevant data is collected. However, as it is important that longitudinal data is available so that trends and changes can be tracked over time, it is also critical that the review process takes this into account. In particular statistical elements should not be deleted without appropriate consideration of their long-term importance. The challenge is then to keep the total number of data elements manageable whilst incorporating new measures and maintaining sufficient data to record trends.

Hence, expanding the title of this paper, library statistics should

- reflect yesterday i.e. record and measure the past
- reflect today i.e. incorporate new data elements whilst retaining continuity with past statistical reports
- reflect tomorrow i.e. develop measures which will measure and record evolving library services.

How will data be collected?

There are many tools and methods, from simple counting (either by a person or using devices such as turnstile counters) to extracting data from databases, log files and library systems. Technology makes much it easier, but a large amount of data may be overwhelming and confusing, unless logical processes are developed and reliably implemented to collect and extract the data. It is important that the guidelines and definitions for the data to be collected are as clear and unambiguous as possible so that each library contributing to a statistical data collection actually records comparable data. My experience is that this is a complex process and that it will generally take several iterations with a range of libraries before a definition can be 'signed off'. Even then it is sometimes surprising how a new staff member or a library reviewing their processes e.g. when upgrading their library system, will interpret the definition in a way that was never envisaged.

Who are the statistics for?

Different statistics may be appropriate for different audiences and at different times, e.g. for library clients, for library stakeholders, for library funders. Libraries do not have any inherent objective value – value is a subjective concept and is generally related to perceptions of actual or potential benefit.

CAUL Statistics

Statistics have been collected annually for the Australian University Library community since 1953, and were originally published in the "News Sheet of the University and College Libraries Section, Library Association of Australia". They are now published annually in print form in the journal AARL (Australian Academic and Research Libraries), and all data back to 1983 has been made available on the CAUL web site as Excel spreadsheets. New Zealand University library data has been

included since 1974. CAVAL has been managing the annual collection and publication of the Australian and New Zealand academic library statistics for CAUL since 1992. During this period CAVAL has also undertaken the retrospective conversion of data from earlier years into Excel format and has provided statistical consultancy services to CAUL, universities and a range of other clients.

The table below summarises the changes in the dimensions of the data collection over 50 years. The number of contributing institutions increased with the creation of new universities and colleges in the 1960's and 1970's, and then reduced following the cycle of mergers and amalgamations which took place in the late 1980's and early 1990's. The number of data elements reached a peak in 1992, this was reduced over the following decade to make the collection more manageable.

	1953	1961	1983	1992	2003
Contributing institutions	9 Australian University libraries	10 Australian University libraries	26 Australian & New Zealand University libraries & 52 College libraries	72 Australian & New Zealand University Libraries - (main, campus & branches)	39 Australian Universities 1 Australian archival store 8 New Zealand Universities
No of data elements	14	84	81	97	85
Data categories	<ul style="list-style-type: none"> • Staff • Holdings • Expenditure 	<ul style="list-style-type: none"> • Staff • Lending • Inter-library loans • Seating • Acquisitions • Holdings • Expenditure • Institutional population 	<ul style="list-style-type: none"> • Library Staff • Administrative structure • Library Services • Bibliographic Resources • Library Expenditure • Institutional population 	<ul style="list-style-type: none"> • Library Organisation • Library Staff • Library Services • Collection Resources • Expenditure • Institutional population 	<ul style="list-style-type: none"> • Library Organisation • Library Staff • Library Services • Collection Resources • Expenditure • Institutional population

The data from 1953 to 1983 was published in the “Red Book’ of Library Statistics (Long, 1986).

Representatives from the various library sectors in Australia met to review the ‘state-of-the-art’ of statistical collections in libraries in 1990 in Perth, Western Australia. The report of this meeting (Exon 1990) laments the lack of national and cross-sectoral cooperation to create reliable and comprehensive statistics that can be used to support libraries. Now, fifteen years later, this has not yet been achieved.

Until 1996 the CAUL Statistics data was collected using printed forms and the data was then keyed into a mainframe SPSS database from which the output reports were generated and printed externally. This process was streamlined in 1996 when the use of Excel spreadsheets and email were introduced. Contributing libraries completed an Excel spreadsheet that incorporated the data definitions and a template for the data values to be entered. The CAUL Statistics website was established in 1997 in order to make the collated statistical data easily accessible and publicly available. Since that time the website has expanded to incorporate the data definitions and Excel templates for each year’s statistical collection, along with links to other relevant statistical websites and the CAUL Statistics reports back to 1993 (those prior to 1996 were converted from SPSS to Excel spreadsheet format).

In 2003 the CAUL Statistics Focus Group conducted a survey of CAUL members to find out how the statistics are used, and requirements for further development of the statistics. There was strong support for a number of enhancements, especially for an online statistics website. The well-known Association of Research Libraries (ARL) online statistics website had long been considered a desirable model. This site was developed and is maintained by the University of Virginia. In 2003 CAVAL established a partnership with the ARL, one aspect of which was the opportunity for CAVAL to work with the ARL and the University of Virginia to develop a pilot CAUL Online Statistical site. This was achieved in mid 2004, and the site was then tested and reviewed by CAUL staff responsible for data collection – their response was enthusiastic. Following the pilot CAUL engaged CAVAL to develop and implement the site as an ongoing service and specified that an online data entry module be incorporated. The site, at <<http://statistics.caul.edu.au>>, went live in early 2005. Data back to 1995 will be accessible via the online site, currently 6 years data is available and the rest of the data will be added over the next few months. The collection of the 2004 data is currently in process using the new online data entry module.

The design and development of the online database raised some complex issues in relation to data relationships and definitions. For example, what degree of change to the name of a data element and/or its associated definition is required before it should be regarded as a different variable? It was a relatively easy decision to make the current item “Non-serial items: total” equivalent to the older “Total monograph items in library” and even older “Total monograph volumes in library” data items. However the former “Most common number of opening hours during semester” and “Most common number of opening hours during vacation” are clearly not equivalent to the current “Opening Hours”. Similar issues apply to institutional names, although it can be argued that a name change is simply a label, whereas a merger or amalgamation creates a new organisation. In the past some (but not all) multi-campus University libraries supplied data for each campus, whereas nowadays data is only collected for the library as a whole. Each of these relationships needed to be clearly defined when mapping the data into the online database. A decision was made to discontinue the use of column numbers in the online system. Although column numbers have been a useful shortcut reference to data elements, they cause confusion when items are added or removed over time.

CAUL Statistics Focus Group

The CAUL Statistics Focus Group is a committee which oversees the collection and publication of the CAUL Statistics. The Terms of Reference and Operating Guidelines state that “the CAUL statistics represent a major achievement of the organisation, and provide a clear, reliable basis for benchmarking, comparison and analysis. They compare favourably with data produced by equivalent bodies in their useability and consistency. They constitute a major resource for the management of university libraries in Australia and New Zealand.”

The goals of the Group are that the CAUL Statistics provide

- Accurate, relevant and authoritative data relating to Australian & New Zealand university libraries
- Consistency of data over time
- Data available widely – accessible, public and widely promoted

- Data which permits analysis of trends
- Collection of both traditional and new information
- Training/feedback opportunities for relevant university library staff
- Statistics which are cheap, useful and valid.

The CAUL Statistics Focus Group works closely with CAVAL to determine the fields for which information is to be collected and develops definitions and instructions. It is the role of the group to ensure that the statistics remain relevant, and capture as much as possible the range of activities and outputs of a university library, so that the development, trial and incorporation of new fields into the collection are regularly discussed. The group also monitors the CAUL statistics ensuring that they are aligned, when appropriate, with other statistical publications, in particular those of ARL, CARL and CURL/SCONUL, as well as international standards, including ISO TC 46 and NISO Z39.7 and the work of the Standards Australia Committee IT/19.

As libraries moved to subscribing to serials in electronic form during the 1990s the collection of statistical data about serials increased in complexity. The completeness and accuracy of the data supplied to CAUL was variable. In 1999, the CAUL statistical definitions were expanded to specifically include several categories of electronic serials and these were further refined in 2000. This has resulted in a discontinuity of data and is a complicating factor when analysing the CAUL serial subscription statistics. The CAUL Statistics Deemed List was developed to assist its members in their annual statistical collections by providing a single source of information for the number of journals in various full text packages, either those from a single publisher, or aggregations of several or hundreds of publishers. The objective is ensure that the same methodology is used by all participating libraries so that appropriate comparisons can be made – rather than being distracted by exactness, for example, whether a particular collection has 543 or 544 titles. The information is collected from publishers, aggregators and web-sites. At this point in time it is acknowledged that it is difficult to de-duplicate titles held in both print and electronic collections.

Current issues being discussed by the CAUL Statistics Focus Group are the collection and reporting of data for e-books and off-shore students.

Contributing libraries are encouraged to supply corrections to data from previous years. Often this occurs when internal processes are reviewed and/or a different staff member takes over responsibility for the statistical data. CAVAL then updates the database and also the full Excel spreadsheet on the CAUL statistics website. These amendments are included in the next report published in AARL.

The CAUL Statistics Focus Group encourages institutions to use the CAUL Statistics as their only statistical database. For this reason a number of optional non-core data elements are included. Core data is defined as bench-markable, required by other bodies or useful for political purposes. It is recommended that non-core (i.e. optional) data is supplied wherever possible, even if only a best estimate, in order to improve the reporting of time series data. Optional data is not reported in the annual CAUL Statistics published report, however it is included in the online database.

Incorporation of data on electronic resources into the CAUL Statistics has concerned the Group for several years now. Following the well-publicised ARL E-metrics trial, local trials were run in New Zealand (2003) and Australia (2004). As the majority of libraries were unable to consistently collect the data for several of the metrics, only one item, Expenditure on E-resources, has been included in the 2004 data collection. As improved methods of data recording and collection, such as COUNTER compliant datasets, are implemented it is expected that additional metrics will be included over the next few years.

Data Quality

Genoni (2004) examined the use of the CAUL statistics for longitudinal analysis of serial subscriptions and described a number of issues that made this difficult. These included incomplete and inconsistent data, changes in methods of collecting or presenting statistics, inaccurate calculations and the addition of new universities. He supports the continuing effort to update and correct the data whenever possible - "It might be argued that any errors in the past CAUL statistics are of little interest. The value for the effort and resources that are put into compiling the statistics in this and future years is, however, derived in large part from what they tell us about the way in which library services and collections evolve to meet their changing environment. This can only be achieved by statistics that can be relied upon for accurate longitudinal analysis. It might be too late to correct some of the uses which have been made of the existing inaccurate statistics, but the statistical record itself should be made as complete and accurate as possible".

Commonly perceived components of data quality are that data should be

- complete, error-free and based on consistent definitions
- reliable, i.e. consistent, dependent, stable and predictable
- valid and accurate.

Considerable progress will be made towards a high quality statistical data collection where staff from contributing institutions have appropriate training and good support from the data collection agency, where there is clear and open communication between everyone involved in the collection and publication processes, and there is strong management support and appropriate internal procedures within each contributing organisation.

Conclusion

It is vital that statistical collections and processes are regularly reviewed and updated in order to ensure that they remain relevant and useable in the changing information environment. The maintenance of a consistent and reliable statistical data collection is an essential but complex and important background task for libraries.

REFERENCES:

ARL (Association of Research Libraries) - <http://www.arl.org/stats/index.html> and
ARL Online statistics - <http://fisher.lib.virginia.edu/arl/>

CAUL (Council of Australian University Librarians) statistics site -
<http://www.caul.edu.au/stats/>

CAVAL/CAUL Online statistics – <http://statistics.caul.edu.au>

E-metrics (ARL) <http://www.arl.org/stats/newmeas/emetrics/>

Exon, F.C.A. & Smith, K. (Editors) (1990), "National Think Tank on Statistics". Papers presented at a meeting held on 29 September, 1990, Perth, Western Australia.

Genoni, P. (2004), "Print Serial Collections in University Libraries Post 1990: What do the CAUL Statistics Reveal?" *AARL*, vol 35 no. 3, pp. 177-192.

ISO (International Standards Organisation - TC 46 – Information and Documentation) -
<http://www.iso.org/iso/en/stdsdevelopment/tc/tclist/TechnicalCommitteeDetailPage.TechnicalCommitteeDetail?COMMID=1757>

Long, A. (Editor) (1986), "Australian and New Zealand Academic Library Statistics Part A: Tables 1953 - 1983" *Western Library Studies* No. 6, The Library, (WAIT) Western Australia Institute of Technology.

NISO Z39.7-2002 Draft Standard for Trial Use Information Services and Use: Metrics & statistics for libraries and information providers--Data Dictionary. National Information Standards Organization (NISO) Created by NISO Standards Committee AY: Library Statistics. <http://www.niso.org/emetrics/>