Quantitative aspects of the management of the modern (scientific) library

QUANTITATIVE ASPECTS OF THE MANAGEMENT OF THE MODERN (SCIENTIFIC) LIBRARY

Leo Egghe
LUC and tUL
Universitaire Campus, B-3590 Diepenbeek, Belgium*
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

Abstract

This paper and talk examines aspects of data collection for the management of a modern (scientific) library. We discuss: reports as a public relations and public awareness tool, norms and standards, data gathering and its problems in an electronic environment, indicators, complete and incomplete data (sampling) and their uses.

I. Introduction

Services must continuously prove themselves: prove that one is still needed and even that one performs better with shrinking budgets. This is the more true for libraries, being services that are not money making but, on the contrary, cost (lots of) money, with users who are, in general, not very "library minded". We must be able to convince subsidising leading bodies that libraries play an increasingly important role in modern scientific life. This "proof" can be given by establishing public relations (PR) and public awareness (PA) activities. By PR is meant: making libraries and their services attractive ("sell" the services) and publicising the library; by PA is meant: informing directors, making them aware of activities and needs of the library and make sure that libraries are in their minds when they have to make (budgettary) decisions.

The physical form of PR and PA actions include diverse reports of several types: for PR: informative brochures, guides, folders, WWW pages in an attractive form, press releases (or press conferences); for PA: special reports prepared for discussions in meetings, annual reports. Of course, in some cases annual reports do also serve as PR tools, especially in the case of company information.

All these PR and PA actions do not only serve the goal of informing the subsidising leading bodies: they are also needed to inform the library users. Giving them good information leads to an understanding of certain rules or to the acceptance of the fact that one is obliged to ask a fee for certain services.

Last but not least, reports are also needed for the library management itself: only by producing professional reports the librarian and his/her staff are able to fully "understand" what is going on in the library (e.g. by revealing many "hidden" activities such as the reshelving of books) and to predict future uses and problems. Reports can also include aspects of benchmarking, i.e. comparisons with comparable libraries.

Reports can only be produced in a professional way by collecting concrete and sufficient "hard" correct **data**. Collecting such data is very well possible in the everyday life of a library but its complexity is sometimes underestimated and is growing in time.

II. Data

II.1. Definition of the investigated property

Investigated properties must be clearly defined and must be the same for all library staff members otherwise they will report differently on the same topics, hence making the outcome useless. Two examples of possible confusions:

"Money spent for books" is ambiguous unless specified: are we interested in the budgeted amount, the amount of money appearing in the orderings, the total price of the delivered books, the total amount of the invoiced books or the total amount paid for books in a certain year? Is the definition of "book" clear (e.g. does it include serials, theses,...)?

"Number of circulations" are continuations included or interlibrary circulations?

A clear definition of the investigated properties is also essential when common (or comparable) reports among similar libaries (e.g. in a library consortium) are to be made (useful in benchmarking). Often many meetings of

---

* Permanent address.

co-workers (or colleagues) are necessary in order to get a common view and understanding of the topics to report on. To reach this goal one can use internationally accepted definitions or standards.

Internationally recognised statistics or norms and guidelines (incl. scope notes) might be of help in this matter. Examples of statistics: LIBECON (EU) (see http://www.libecon2000.org), ARL-Statistics (ARL=Association of Research Libraries) (USA) (see http://www.arl.org), UNESCO (2000). Norms and guidelines: Ward, Sumsion, Fuegi and Bloor (1995) (EU guidelines on which the BULLS statistics, i.e. the statistics of the Belgian university libraries, are based - see http://www.ua.ac.be/BULLS/), Poll and te Boekhorst (1996) (IFLA guidelines), ISO (1991) (the ISO 2789 guidelines on library statistics of UNESCO).

II.2. Organisation of data gathering

Data are usually collected on a yearly basis, e.g. for the annual report. Collecting data needs a continuous effort throughout the year. Only when data will be generated by a computer, they can be produced in the beginning of the next year (e.g. number of circulations, number of books catalogued). In all other cases a kind of "logbook" (e.g. for counting the number of books reshelved or for counting the number of library visitors) or a file (e.g. of invoices) must be kept. From the above it is clear that, along with the discussions on (further) automating the library, the issue of the automatic generation of library statistics must be kept in mind (see also section II.3).

It depends on the type of library and even of the local situation which topics are important to report on to prove something or to support certain claims, see Egghe and Rousseau (2001) or LUC (2002) for some proposals. It is better to collect few data in an accurate way than many inaccurately or unreliably.

Very important is the "added value" when **the same** data are collected over a long time period (i.e. several years). Then "derived results" such as time analyses (e.g. regression lines predicting trends in time) can be produced. It is therefore strongly advisable, once an agreement on the definition of a topic is reached, to keep this definition fixed. Otherwise time series are worthless. Notwithstanding, of course, the need to introduce new statistical data e.g. on new technologies (e.g. use of electronic journals).

The building of a good data collection takes a long time. It is clear that if one decides in year **x** to report on a new topic, one can only start collecting in year **x+1**, hence reporting on this topic can only start in year **x+2**. Add to this several more years in order to have a time dimension in these data!

II.3. Data gathering in an electronic environment

In recent years, the number of "electronic" activities has increased drastically. In most cases this also means that data are gathered in an automatic way and hence there may be a perception that it has become easier to collect data. This is **not** true. It is true that data are gathered more quickly, but at the same time their accuracy has dropped. One reason can be the fact that these data are delivered by the computer (server) via a third person who might have another insight on the exact definition of a certain attribute. This is even more true when libraries are collaborating in an electronic network. In that case it may happen that quantitative topics wanted by different network partners are not exactly the same and also different from the corresponding ones delivered by the network's computers (and sometimes one even does not know the difference!).

An example is the information on users (based on barcodes): are all users counted, or only the ones that were activated this academic year (i.e. the ones who used the library at least once this academic year). Another problem is reporting on the number of books added to the collection this year: are free books included (e.g. theses), are new editions included, are multiple copies included, how are serials counted, and so on? The problems arise because the data are generated by a computer (and not by each librarian manually) and hence it is not easy to make sure that what is in the librarian's mind is also delivered.

Another reason for the increased problem of automatic data gathering is that during the year there will be some periods of system breakdown and subsequent loss of data. Sometimes this is not observed, sometimes it is and a method of "interpolation" or "extrapolation" is applied, but in any case the final result is not exact. An example is given by unregistered circulations of books.

The problem of data gathering in an electronic environment has been complicated by the increase of web-oriented library activities. A typical example is a web OPAC. Let us give an example from the library of Limburgs Universitair Centrum (LUC), belonging to the Anet network. The library catalogue was automated in 1989 and became a web catalogue in 1995. Between 1989 and 1995, we were able to report on the search time in the library's OPAC. This is no longer possible for the web OPAC. A similar problem is experienced by DIALOG users. Users accessing DIALOG via the WWW find that, instead of connect time being indicated (and invoiced in this way), DIALOG units are used instead and they cannot be used to measure connect time in a file.

We have come across a major difference between the internet (the virtual world) and the real world: in the latter, "use" is measured by **time**; in the former "use" is measured by **number of times**. It is not clear what the impact of such a big change will be on the (social) habits of information exchange.

Even "number of times" is sometimes difficult to measure. Let us go back to the example of the LUC OPAC. Since it has become a web OPAC, contact is possible from outside the library and even from any place in the world. It is therefore

- not easy to report on the number of OPAC contacts
- not very relevant to report on all these contacts since an OPAC search, from India, for example, to the LUC catalogue has a different goal than an OPAC search within the LUC library: in the former the OPAC is used as a documentary system in which an information retrieval (IR) process is going on; in the latter the OPAC is often used as a library catalogue.
- Mixing these OPAC searches does not make much sense but separating these different OPAC uses is not possible since, in some (or even many) cases, OPAC searches within the library have an IR goal and are not used to obtain knowledge of the library's holdings. This is even more the case for a web OPAC search in a university library, performed from a professor's office!

More and more, OPACs serve for IR purposes, thereby partially replacing searches in subject-oriented databases (such as Chemical Abstracts, Inspec, Econlit,...) often offered by commercial hosts (such as DIALOG). Also the WWW is used as an alternative for these (relatively) expensive IR services. Together with this evolution, there is a move away from services executed by professional library staff to actions performed by untrained users. In addition to this there is the problem of the enormous size of the internet (and of the WWW) and its fast growth (for example see Egghe (2000)). We conjecture that all this implies lower quality of the IR processes (for example in the sense of recall and precision - see again Egghe (2000)) although we must admit that it is extremely difficult to express this in a quantitative way. One major reason for this is that, when searching in the WWW, one does not obtain a clear set of documents but a ranked, truncated list, ranked according to the expected relevance for the searcher. So one can no longer report on "number of documents" retrieved: even "retrieved" becomes a fuzzy notion since, in a search result of say 5000 documents, the first ones (as presented in the ordered truncated list) are "more retrieved" than the ones on ranks near 100 and these in turn are "more retrieved" than the ones on ranks near 1000 or 5000 (these ones are probably not retrieved in the sense that they are not used). We refer to Egghe and Michel (2002a,b) for first attempts to measure similarities in ranked outputs.

Another problem with web information is that it is usually not dated and often anonymous. This makes it difficult to report here on number of authors, obsolescence of the literature, and so on. Because of the lack of time information, the hyperlinks (clickable buttons) in web pages cannot be considered as the web analogue of classical references or citations (see also Almind and Ingwersen (1997), in which the notion "publication time" is replaced by the notion "real time").

New, accurate definitions are in order, probably to be formulated as ISO standards by the International Organisation for Standardisation cf. ISO (1991), for this virtual (but also very real!) world. We must adapt and accept that classical data types (such as connect time) have to be replaced by new data types (such as number of connections).

III. Indicators

The term "indicator" describes a wide variety of "derived data" and constitute the first stadium of "data crunching" after the data gathering itself. Usually, an indicator is defined as the division of two data, usually representing a fraction.

Examples:

- The number of satisfied outgoing interlibrary requests divided by the total number of outgoing interlibrary requests, i.e. the fraction of satisfied outgoing interlibrary requests, i.e. the "success-ratio".
- The number of books in the French language divided by the total number of books in a library, i.e. the fraction of books written in French.
- The fraction of mathematics books in a library.

A fraction multiplied by 100 is a percentage. An example of an indicator that is not a fraction is given by the division of the number of books in French by the number of books in English, indicating their relative sizes. Also

important is the division of the number of incoming interlibrary requests by the number of outgoing interlibrary requests. Another indicator that is not a fraction is given by growth or aging rates, i.e. numbers of the form

$$\frac{f(t+1)}{f(t)}$$

where t=time (e.g. a year) and f(t) e.g. denotes the number of books purchased by a library in year t (t+1 being one year later - example of growth) or f(t) denotes the number of references (in e.g. a journal) to t years ago (t+1 being one year earlier - example of (synchronous) aging). Finally, the impact factor (in general "number of citations to a journal divided by number of articles in this journal") is also a well-known example of an indicator that is not a fraction. For more examples, see Lafouge, Le Coadic and Michel (2002). Fractions are - however - mathematically treatable e.g. in the connections of the treatment of statistical samples (see further).

IV. Complete and incomplete data

From a statistical point of view, complete data do not exist. All measurements or data gathering activities yield a moment's vision or constitute a sample from a much larger population, such as a library. Nevertheless, for purposes of library management, it is convenient to make a distinction between complete and incomplete data sets.

Complete data are obtained if one wants to report exactly on what one measures. For example:

- The number of borrowings in a year is obviously a complete result. We do not intend to say anything about the circulation behaviour in libraries elsewhere in the world.
- The price of the books that are purchased in a certain year: these are complete data if we only want to report on this. We cannot, however, conclude anything on the price of books elsewhere (e.g. worldwide or at a bookseller). If we want to do this, we have incomplete data (so called sampled data).
- The number of times that the OPAC has been used in a year. Since we only want to report on the OPAC use in our library, these data are complete.

Most policy decisions are based on incomplete data, that is on samples. Indeed, evaluations and managerial decisions can be made only when there is a vision on the totality in the whole library (or even the world). To know this totality is usually an impossible requirement and one is led to make total conclusions based on a relatively small sample. Some examples:

- The difference in average delivery times between two booksellers can only be estimated on the delivery of books to your library, not on the totality of actions of both booksellers (and often it will even be necessary to sample the books delivered to the library).
- Are there more co-authors (averaged per book) in chemistry than in mathematics? Obviously, only a sample can be taken, certainly when one wants to answer this question in a worldwide vision but probably also within the library since the number of these books might be quite high.
- Asking users' opinions on the library services or asking the population (including non-users) on some library issues will obviously be limited to samples.

In conclusion, one must determine the "universe" Ω: this is the total population (of persons, books, etc.) on which one wants to measure a certain characteristic (cf. the above examples). The size of Ω then determines whether or not we have to draw a sample. Indeed, the size of Ω determines the time (hence the money) we would have to spend on collecting a complete data set. If this is possible, it is clear what to do: we have to measure the characteristics on every element of Ω. If sampling is needed, a lot of time is saved but a new problem arises: how to sample?

V. The collection of incomplete data: sampling

How to collect incomplete data is a nontrivial problem. By the very definition of "incompleteness", we are faced with the problem: which elements of the population will be sampled? The main problem is the possible bias, the unequal chance for elements of the population to be picked.

Examples of bias: to give a questionnaire only to library users that are in the library (in this way you might miss the possibly unsatisfied absent users), picking books in the library shelf by measuring lengths (this way thicker books have a higher chance to be picked), checking the length of services by sampling say every 30 minutes, and so on.

A perfect method to sample correctly is random sampling. Every element of the population has an equal chance

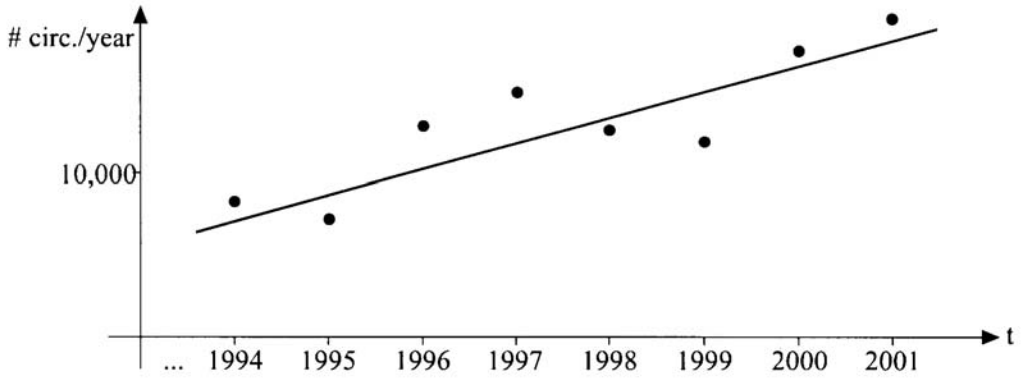

Figure 1: Scatterplot: Number of circulations/year in function of the year and trend analysis

to be picked. The method uses a computer-generated list of (pseudo) random numbers and they determine the elements of the population that should be picked (for materials that are in the computer this random sampling can be executed in an automatic way). Executing random sampling in a manual way is time consuming; faster methods exist.

Systematic sampling samples by length or by time but yields a bias, as remarked above. This can be avoided by applying the Fussler sampling technique. For the example of picking books from a library shelve, this goes as follows. Measure a fixed length (say a meter) but do not pick the book that is indicated this way but take the next one (see Fussler and Simon (1961)). One can prove that, if books are on the shelve in a random position according to thickness, this method is as good as the random sampling technique (and, obviously, is as fast as the systematic sampling technique) (see Egghe (1988)). This technique can also be used for measuring service times at an information or circulation desk.

VI. The use of complete and incomplete data.

VI.1. The use of complete data

If it is possible to investigate **every** member of this population our conclusions are 100% sure, but of course limited to this population. These data are summarised in graphs, such as bar diagrams (for discrete data) and histograms (for continuous data). Examples of discrete data are: the age (in years) of a book or the number of years that a book is in the library, number of authors of medical books, number of library users and so on. Examples of continuous data are: the thickness of medical books (e.g. in cm), retrieval times (in minutes or seconds) in a documentary system, delivery times (in days) of books that are ordered at a bookseller. Note that the divison between discrete data and continuous data is sometimes vague: delivery times are discrete data (i.e. entire numbers of days) but we are only interested in them in groups, say periods of 5 or 10 days in order to estimate the order of magnitude. The same is true for the number of times one connects with a Web-OPAC. Ages in years are clearly discrete while retrieval times in seconds are clearly continuous!

In reports the time dimension is very important. In order to visualise the "trend" in time we can calculate the so-called regression line: on the graph consisting of the different points (with the x-axis representing time t) we add a straight line, fitting the "cloud of points" in the best way. By doing so we can see the trend by visual inspection of the line and are able to make (short term) predictions (see Fig. 1). The question of "how to make or not to make" graphs is dealt with in the literature (see Cleveland (1985)).

"Summary statistics" are also needed to interpret data in a professional way. The basic ones are: the mean (average) denoted by $\bar{x}$, the standard deviation (dispersion) denoted by s and the so-called percentiles. The square of s is the variance.

The average gives an overall view of the data, e.g. the average price of books. The standard deviation and the percentiles yield information about the degree of irregularity of the data, the former (s) is mainly used with incomplete data, the latter (percentiles) with complete data. For instance answers can be given to questions such as: how long does it take to deliver 50 or 75% of the books ordered at a certain bookseller?

VI.2. The use of incomplete data

VI.2.1 The case of one data set

This case will be illustrated by an example. Suppose we sampled 100 books, delivered by a bookseller and the average delivery time was calculated (say 61 days). When making a complaint and requesting a faster delivery, the bookseller can question the validity of our data because we sampled only 100 books out of 2,000 deliveries. An estimate of the real average delivery time of this bookseller can be made by calculating the confidence interval as follows. With a sample size N (100 in our example), an average $\bar{x}$ and standard deviation s, we are 95% sure that the real "overall" average delivery of this bookseller is in the interval.

$$\left[\bar{x} - 1.96\frac{s}{\sqrt{N-1}}; \bar{x} + 1.96\frac{s}{\sqrt{N-1}}\right]$$

Example: N=100, $\bar{x}$=61 days and s=18 days. Then the real average delivery time is between the values 57.5 and 64.5 days (with 95% certainty).

Other applications could be: thickness of medical books, length of reference lists of medical articles, number of authors of medical books, or, in the area of user studies: fraction of certain types of users of the library, fraction of users who agree with a change of the opening hours, fraction of the loan transactions containing the maximum number of books that are allowed to check out at one occasion, fraction of books that are returned too late, fraction of lost/stolen books, fraction of users that are interested in an SDI-service on the acquisition of the library. Finally we mention another important application: measuring the overlap between two libraries or between two databases: O(B/A)=the overlap of database B w.r.t. A (i.e. the fraction of the articles in A that are also in B).

Note that fractions are averages and hence for fractions one can calculate confidence intervals as explained above. That fractions are averages can be seen as follows: the fraction of a subset A in a universe Ω is the average of the numbers

$$\{\chi_A(\omega) \mid \omega \in \Omega\}$$

where χ_A is the characteristic function of A, i.e. $\chi_A(\omega) = 1$ if $\omega \in A$ and $\chi_A(\omega) = 0$ if $\omega \notin A$.

VI.2.2 The case of two data sets

It is an interesting issue to compare two populations. For example to compare two libraries w.r.t. to their average speed of delivery of interlibrary loan requests. As we will have here two incomplete data sets, what can we say about the average difference? Techniques in statistics allow us to draw e.g. 95% or 99% conclusions about the possible different behaviour.

Examples of application: difference between the average delivery times for books at two booksellers or of books coming from different countries (or, as mentioned above, of interlibrary loan material coming from different libraries), difference between the average thickness of medical books and of mathematics books, difference between their average number of authors, fraction of female users of library A versus the same in library B. Final general example: difference of use when we consider two different time periods (measuring changes in quality or in use of certain services).

For more information on the topics of this talk and paper and for more on informetrics-bibliometrics-scientometrics we refer to the books Egghe and Rousseau (1990, 2001). A more complete list of library statistics of a scientific library can be found in Egghe and Rousseau (2001) and LUC (2002).

References

Almind, T.C. and Ingwersen, P. (1997). Informetric analyses of the world wide web: methodological approaches to "webometrics". Journal of Documentation, 53, 404-426.

ARL: see: http://www.arl.org/

BULLS: see: http://www.ua.ac.be/BULLS/

Cleveland, W.S. (1985). The Elements of graphing Data. Wadsworth, Monterey (CA), USA.

Egghe, L. (1988). The Fussler sampling technique for populations with a discrete or a continuous distribution of thicknesses. In: Informetrics 87/88 (L. Egghe and R. Rousseau (eds.)). Proceedings of the first international Conference on Bibliometrics and theoretical Aspects of Information Retrieval (LUC, Diepenbeek, Belgium, 1987), 65-74.

Egghe, L. (2000). New informetric aspects of the Internet: some reflections - many problems. Journal of Information Science, 26, 329-335.

Egghe, L. and Michel, C. (2002a). Strong similarity measures for ordered sets of documents in information retrieval. Information Processing and Management, 38(6), 823-848.

Egghe, L. and Michel, C. (2002b). Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. Information Processing and Management, to appear.

Egghe, L. and Rousseau, R. (1990). Introduction to Informetrics. Quantitative Methods in Library, Documention and Information Science. Elsevier, Amsterdam. ISBN 0-444-88493-9.

Egghe, L. and Rousseau, R. (2001). Elementary Statistics for effective Library and Information Service Management. ASLIB-IMI, London (UK). ISBN 0-85142-451-1.

Fussler, H. and Simon, J. (1961). Patterns in the Use of Books in large Research Libraries. University of Chicago Press, Chicago (USA).

ISO (1991). Norme Internationale ISO 2789. Information et Documentation - Statistiques internationales des Bibliothèques, ISO, Genève.

Lafouge, T., Le Coadic, Y.-F. and Michel, C. (Préface de Egghe, L.) (2002). Eléments de Statistique et de Mathématique de l'Information. Infométrie, Bibliométrie, Médiamétrie, Scientométrie, Muséométrie, Webométrie. Presses de l'enssib, Villeurbanne (France).

LIBECON: see: http://www.libecon2000.org

LUC (2002). Jaarverslag Universiteitsbibliotheek LUC 2001, LUC, Diepenbeek.

Poll, R. and te Boekhorst, P. (1996). Measuring Quality. International Guidelines for Performance Measurement in academic Libraries. IFLA Publications 76, K.G. Saur, München.

UNESCO (2000). UNESCO-Questionnaire sur les Statistiques relatives aux Bibliothèques. UNESCO, Paris.

Ward, S., Sumsion, J., Fuegi, D. and Bloor, I. (1995). Library Performance Indicators and Library Management Tools. European Commission, DGXIII-E3, Luxemburg.