COMMENT

OBC
www.rsc.org/obc

Chemistry on the world-wide-web: a ten year experiment†

Jonathan M. Goodman
Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge, UK CB2 1EW. E-mail: J.M.Goodman@ch.cam.ac.uk; Fax: +44 1223 336362; Tel: +44 1223 336434

Received 1st July 2004, Accepted 2nd August 2004
First published as an Advance Article on the web 3rd September 2004

The server logs for access to the Cambridge Chemistry webserver show how use of the server has increased over the last ten years, with access doubling every year and a half. This growth has started to slow, and extrapolation of the data suggests that the current rate of access is close to a plateau of ten million downloads a year. The transition for chemists from no internet access to saturation coverage, therefore, appears almost complete.

Introduction

The Chemistry Department of Cambridge University started a WWW server (http://www.ch.cam.ac.uk/) in April 1994. Since June 1994, weekly access logs have been collected, and there are now ten years of data on access to the web server. These data show how access has grown, since the time in 1994 when there were about a dozen chemistry departments with web servers world-wide, to now, when there are about two thousand.[1] The web server has information about the chemistry department, its academic staff, research and teaching. It also has databases (C2K: http://www.ch.cam.ac.uk/c2k/) and resources, including methods for chemical calculations.[2] Although the web server has been continuously updated over the last decade, its role has not changed dramatically, as many new information resources have become available on separate webservers in the department which are indexed from http://www.ch.cam.ac.uk/ but with access logs that are not included in this analysis.

Studying the whole of the WWW is a near-impossible task, and its size is best estimated by random sampling.[3] This approach has been used to investigate the size of the WWW, and showed there were about 2.8 million publicly accessible web servers in February 1999.[4] Soon afterwards, it was shown that the apparently random growth of the WWW has led to regularities. For example, Huberman and Adamic sorted websites by their number of pages, and showed that the distribution follows a power law.[5] The paper begins by referring to the 'exponential growth' of the WWW, but further data suggest that the overall growth is not exponential. The evolution of the WWW as a whole has been studied by the OCLC.[6] The number of websites has risen dramatically (Fig. 1), but the graph suggests that this will not continue indefinitely. The number of websites available worldwide was nearing nine million in 2002, and three million for the public WWW, defined as those websites which offer free, unrestricted, access to a significant proportion of their content. Whilst the number of public websites decreased slightly between 2001 and 2002, the average size of each public site increased from 413 pages to 441. This suggests that the public WWW increased slightly, but not dramatically, between 2001 and 2002.

These studies were based on random sampling, and so there is a degree of uncertainty in the figures. The access logs for the University of Cambridge WWW server are much more precise, but measure only one small corner of the internet. The logs appear to show dramatic and continuing growth in access to the web server, but this cannot go on forever, as it is unlikely that everyone on earth will want to download files from this web server every week. At some point, this rapid growth is likely to slow. The rapid growth is due not only to the growth of information on the web server, but also to the increase in internet use by chemists. Ten years ago, many computers were not internet connected. It is now unusual for a computer not to be connected to the internet, and so access to the internet may be beginning to approach saturation point. The analysis of the Cambridge Chemistry log files may help us to discover how close chemistry is to this point.

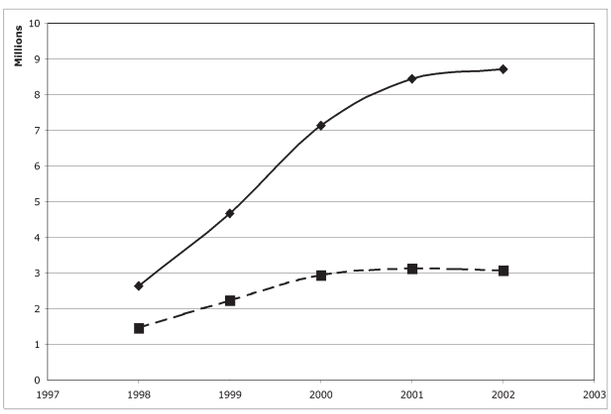

Fig. 1 The total number of websites (data from the OCLC[2b]) for the public WWW (dashed line) and the whole WWW (solid line).

Results

The numbers recorded in the log file are the total number of files downloaded from the web server each week, including HTML files, images, Java programs, *etc.* The general style of the webserver has not changed dramatically over the last ten years. If the server were redesigned so that accessing the home page downloaded dozens of images instead of a total of two or three files, then the weekly access numbers would probably show a corresponding jump. The style of the home page and the rest of the server have deliberately been kept simple, so this effect should not have a major influence on the data.

The logs cover short periods when power failures, network problems or system issues have prevented the server running properly. These periods have been so short that they do not show up clearly in the logs as major dips, although they lead to occasional sharp fluctuations. The four weeks from September 15th to October 13th, 2002 do not have data available, and this has been addressed by using the average of the previous week and the following week for this period. The same has been done for November 30th to December 14th, 2003. With these exceptions, the numbers show the activity of the server over the last decade (Fig. 2), and the graph shows growth

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

DOI: 10.1039/b409956g

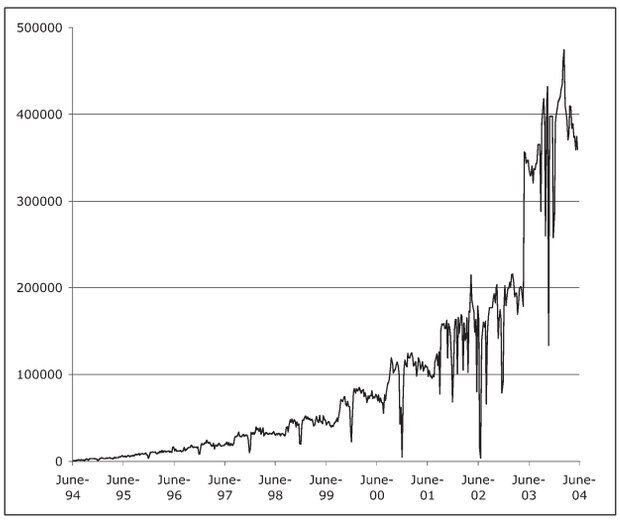

Fig. 2 Weekly WWW access: http://www.ch.cam.ac.uk/.

that is approximately exponential ($r^2 = 0.92$, $y = 3800e^{0.0092x}$; x represents the number of the week, starting at $x = 1$ in June 1994; y is the number of files downloaded in that week).

Discussion

In 1996, an effort was made to predict the growth of access to the webserver (Fig. 3). At this stage, linear growth had a very similar correlation coefficient ($r^2 = 0.91$; $y = 111x + 29$) to exponential growth (transformed regression model: $r^2 = 0.89$; $y = 1443e^{0.0225x}$; untransformed regression model: $r^2 = 0.91$, $y = 1976e^{0.0184x}$; the Experimental section gives information on the models used).

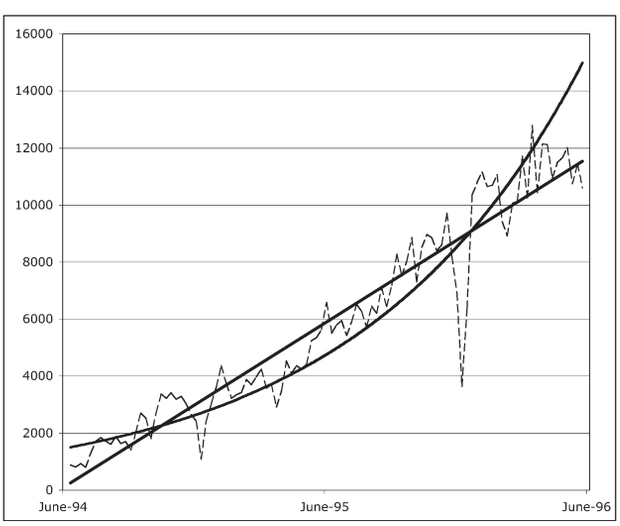

Fig. 3 Linear and exponential regression models for the first two years.

Linear extrapolation for the next year suggested that seventeen thousand hits a week should be expected in June 1997, and untransformed exponential extrapolation suggested almost thirty five thousand a week. The actual figure was around twenty thousand. Linear extrapolation suggests that access should be running at almost sixty thousand per week at the end of the first decade. Exponential extrapolation suggests a rate of twenty-eight million a week. The actual figure of around four hundred thousand per week falls within this rather generous range.

When applied to the first ten years of data, the transformed exponential regression model gave a better result ($r^2 = 0.92$, $y = 3800e^{0.0092x}$) than the untransformed model ($r^2 = 0.74$), particularly for the earlier years, where the untransformed fit minimises the absolute difference between the equation and the data and the transformed fit minimises the fractional difference. Linear regression did not fit the data well ($r^2 = 0.70$).

The graph in Fig. 2 shows a clear fluctuation over each twelve month period. In order to consider this annual variation, each data point was divided by the corresponding value of the exponential fit, and the ten years were plotted. Fig. 4 shows the results. The solid line is the average annual variations for just the first nine years, as the December data is missing for the tenth year. The dotted lines show the results for each of the ten years. The graph runs from July to June, and the most obvious feature is the dip at Christmas and New Year. Also pronounced are changes corresponding roughly to university terms, with a rise in October, and a fall in April, perhaps as students begin to focus on revision rather than exploration.

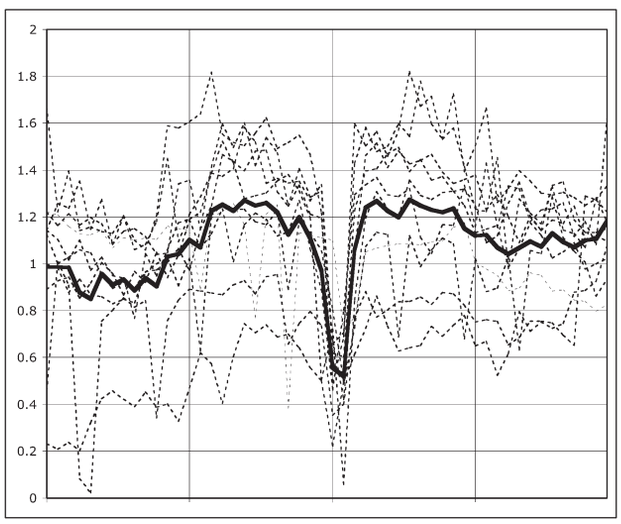

Fig. 4 Annual variation, from July to June.

The average annual pattern was scaled so that it had a mean of one. Each point on the original data set was now divided by the corresponding point on the average annual pattern. This procedure generated a smoothed line without the seasonal fluctuation. This produced a sharp spike in December 2003, due to the use of average data at this point. These three entries were replaced with the average of the previous and subsequent point. This revised curve (Fig. 5) fits to an exponential line $y = 3900e^{0.0092x}$ with $r^2 = 0.93$. This is a slightly better fit than before, with very similar parameters.

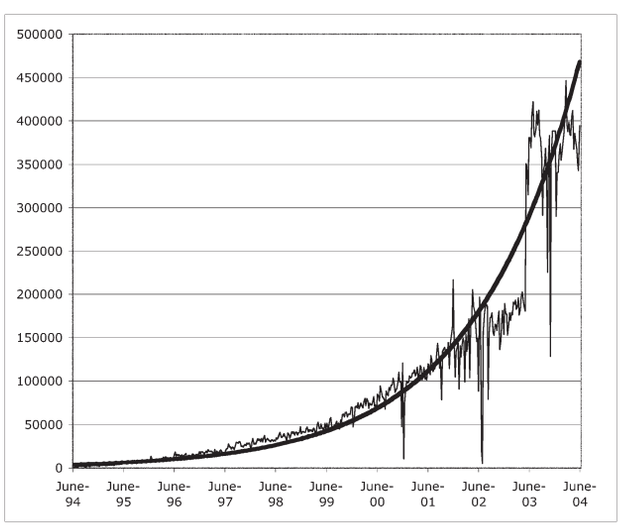

Fig. 5 Data with seasonal fluctuations removed.

This new graph shows a pronounced discontinuity during May 2003, when the average access jumps sharply upwards. The reason for this turns out to be the introduction of a new resource on the web server, http://www.ch.cam.ac.uk/display/ that provides a continuous display of information through a web page, which is regularly refreshed to allow for updates. These

web pages are available anywhere in the world. The display runs continuously in the Cambridge Chemistry Department. The effect of this can be subtracted from the data, to create a rather smoother curve, Fig. 6. This led to one negative value that was replaced with the result for the previous week. ($y = 4353e^{0.0085x}$; $r^2 = 0.91$).

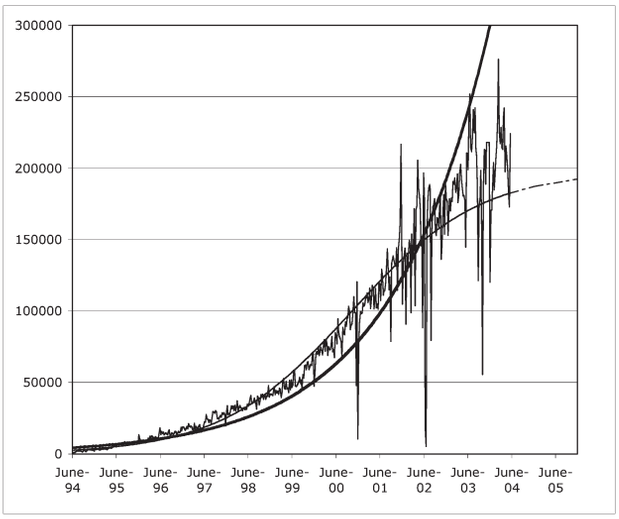

Fig. 6 Smoothed results with exponential fit and eqn. (1).

All of these graphs give a reasonable fit to exponential growth, with access doubling in about eighty weeks in all cases (for the equation $y = Ae^{Bx}$, y doubles when x increases by $\ln 2/B$).

It is not possible for access to continue exponentially without limit. At some point, computer use will saturate, and the curve should be expected to level off. Instead of fitting to an exponential function, this behaviour would be better modelled by a function such as eqn. (1):

$$y = \frac{A}{1+\exp\left(\frac{B-x}{C}\right)} \quad (1)$$

For this equation, y tends to A as x gets very large, to zero as x becomes very negative, and has the value of half of A when $x = B$. It resembles exponential growth for small values of x. This equation was fitted to the data in Fig. 6, following the transformed regression model for the exponential fit, minimising the sum of the squares of the difference of logarithms of the function and observed data, in order to be consistent with the tranformed exponential fit, and in order to ensure that the smaller values were not dominated by the larger ones.

The sum of the squared log difference for the new line is about half that for the exponential growth line. The parameters for the fitted equation are: $A = 200\,000$; $B = 332$; $C = 77.1$. This may be interpreted as suggesting the curve will plateau at about 200 000 downloads a week, and the halfway point was reached by the end of 2000. Since there are about two thousand chemistry departments worldwide (C2K: http://www.ch.cam.ac.uk/c2k/) this represents a hundred for each, on average. An order of magnitude more than this might be surprisingly high, unless a dramatically different resource were to become available.

The same analysis for the raw data (Fig. 2) gives: $A = 400\,000$; $B = 435$; $C = 88.8$. This suggests that the halfway point occurred by 2003, and this is so close to the end of the data that little confidence can be placed on the number. The smoothed results, however, provide strong evidence for the flattening of the curve, even though the data is noisy and the precision of the result may be quite low.

It is possible that this tailing off is due not to a saturation in the number of people looking for chemical information but to the proliferation of chemistry web servers, and so this one web server provides a diminishing fraction of chemistry on the WWW. This was investigated by considering the increase in the number of university chemistry departments indexed in the directory C2K (http://www.ch.cam.ac.uk/c2k/), the number of countries for which chemistry departments are indexed (Fig. 7) and by the size of the Google and the Yahoo directory for chemistry (Fig. 8). Data for the latter were found using the internet archive.[7] All of these measures suggest that the increase in chemistry on the WWW is flattening out. The C2K directory of chemistry departments has entries for 129 countries. There are 191 member states of the United Nations, so completion may be in sight. The number of entries in the Yahoo chemistry directory has increased by less than 1% over the last two years. The Google chemistry directory continues to grow, but less in the last year than in each of the previous two years.

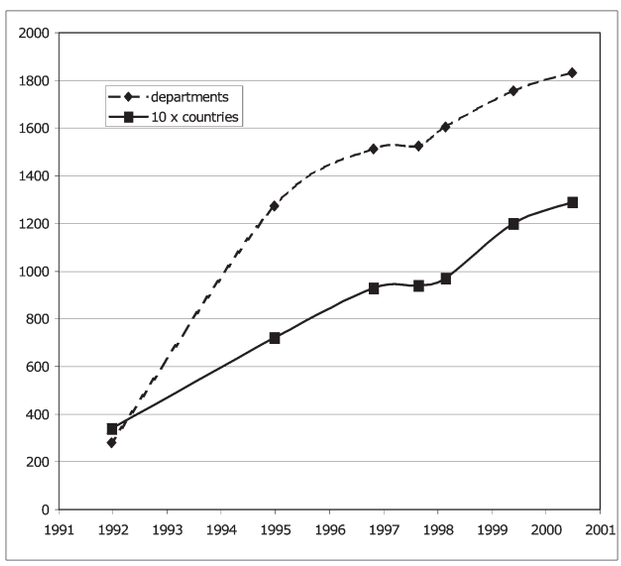

Fig. 7 Number of chemistry departments (dashed line) and number of countries (solid line—the data were multiplied by ten to fit on the same scale) listed in C2K: http://www.ch.cam.ac.uk/c2k/.

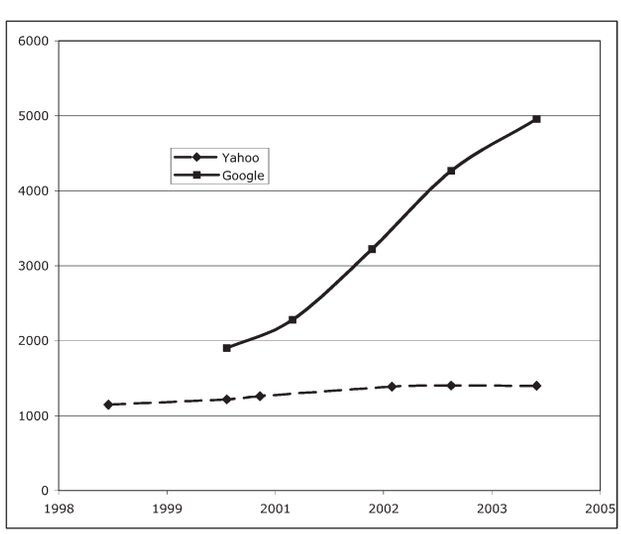

Fig. 8 The number of entries for chemistry in the Yahoo directory (dashed line: http://dir.yahoo.com/Science/Chemistry/) and the Google directory (solid line: http://gle.com/Top/Science/Chemistry/).

The number of web pages that the Google search engine covers are shown on the Google front page (http://www.google.com/). It is currently 4.3 billion, a year ago it was 3.1 billion, and the year before it was 2.1 billion. In July 2001, it was 1.3 billion. This is steady rather than exponential growth. Scientific subjects such as chemistry expanded quickly when the WWW began, before the large increase in commercial sites, and so might be expected to reach saturation before the WWW as a whole.

Conclusions

The impact of the world wide web on chemistry has been dramatic, but these data suggest that the major impact has now passed, and further changes will come not from a continuing increase in the use of the resources which have been available for years but from new ideas. There are many possibilities for the next big advance, perhaps from the new technologies being developed by many groups including the World Wide Web Consortium,[8] or from the innovative use of technology such as RSS.[9] There are many opportunities for exploiting existing ideas, which have yet to be fully explored.[10] For example, molecular structures contain a wealth of information, which cannot easily be searched over the WWW. The technology to do this is available, but it will only be effective if standards are widely adopted. The power of the internet to exploit molecules has yet to be fully realised.[11]

Experimental

A web server was set up on a SGI Indigo R3000 workstation, which ran continuously and had no difficulty dealing with all information requests (up to 100 000 per week), until a disk failure at the end of the year 2000 led to its replacement by a Linux computer. The computers ran the *Apache* web server (http://www.apache.org/). The web server was monitored over a ten-year period (June 1993 to June 2004), and the total files served in each week (Sunday morning to Sunday morning) was recorded.

The data has been fitted using a least-squares approach to exponential growth: $y = Ae^{Bx}$ where x represents the number of the week, starting at one in June 1994 and y the number of files downloaded in that week. The data has also been fitted to eqn. (1). Microsoft Excel uses a transformed regression model, which minimises the sum of squares of the differences between the logarithms of the data and the equation. This was used, and also an untransformed approach, minimising the sum of squares of the differences of the data and the equation. The untransformed approach was less satisfactory, as it biased the fit towards the later years where the numbers are much larger. The transformed approach was used, therefore, both for the exponential fit and for eqn. (1).

Acknowledgements

The referees are thanked for their helpful suggestions.

References

1 J. M. Goodman, *Molecules*, 2000, **5**, 33.
2 (*a*) C. R. Stewart and J. M. Goodman, *Chem. Commun.*, 2003, 2654; (*b*) J. M. Goodman, P. D. Kirby and L. O. Haustedt, *Tetrahedron Lett.*, 2000, **41**, 9879.
3 M. Henzinger and S. Lawrence, *Proc. Natl. Acad. Sci.*, 2004, **101**, 5186.
4 S. Lawrence and C. Lee Giles, *Nature*, 1999, **400**, 107.
5 B. A. Huberman and L. A. Adamic, *Nature*, 1999, **401**, 131.
6 E. T. O'Neil, B. F. Lavoie and R. Bennett, *D-Lib Mag.*, 2003, **9**(4); http://wcp.oclc.org/.
7 Internet Archive: http://www.archive.org/.
8 World Wide Web Consortium: http://www.w3c.org/.
9 P. Murray-Rust, H. S. Rzepa, M. J. Williamson and E. L. Willighagen, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 462.
10 P. Murray-Rust, H. S. Rzepa, M. Wright and S. Zara, *Chem. Commun.*, 2000, 1471; P. Murray-Rust, H. S. Rzepa and B. J. Whitaker, *Chem. Soc. Rev.*, 1997, **26**, 1.
11 H. S. Rzepa, B. J. Whitaker and M. J. Winter, *Chem. Commun.*, 1994, 1907.