# If we do not understand our users, we will certainly fail

***DAVID NICHOLAS***

Director, School of Library, Archive and Information Studies and UCL Centre for Publishing

University College London

*The E-Resources Management Handbook – UKSG*

**This chapter puts forward strategic reasons why librarians should monitor the behaviour of their users, and provides a description of how this might be done, employing deep log analysis methods. In order to illustrate the benefits of doing so, it presents a detailed analysis of virtual scholars and their information-seeking behaviour. The key characteristics outlined are: popularity, diversity, reading/viewing, digital visibility, trust and outcomes. It is concluded that unless libraries switch resources to monitoring and evaluating their users, they risk becoming estranged from their customer base.**

## Introduction

This is a research-led paper which is based on more than five years of evaluating a whole range of e-journal and e-book scholarly information platforms using an evidence-based methodology called 'deep log analysis', a more sophisticated form of transactional log analysis[1]. This paper argues that unless we take on board the characteristics of the virtual scholar uncovered by this work and adopt the kind of evaluation techniques developed during these studies, we shall fail miserably.

Firstly, an explanation is provided as to why we need to obtain a much better understanding of our users; secondly, the benefits that deep log analysis provides in this regard are described; thirdly, in order to whet the appetite and to show the scale of the challenge for information professionals, the key information-seeking characteristics of the virtual scholar identified by The Centre for Information Behaviour and the Evaluation of Research (CIBER) studies are outlined.

## We might be failing

It has to be said that many librarians do not have sufficient understanding of their users and, as a direct consequence, are facing serious problems. Things appear to be coming to a head, and amongst the library community there is a palpable sense of fear (of the unknown) in the air. Indeed, it would not be an exaggeration to say that there is a big risk of libraries being decoupled from their customer base as users migrate in number from the physical to the digital library, the former the province of the librarian and the latter now the province of the publisher. In the increasingly digital information environment, libraries are becoming increasingly anonymous (and unacknowledged) third parties, as users work ever more remotely and directly. The arrival of e-books will undoubtedly accelerate this process as the vast body of students join an exodus, which will also include a good number of social scientists and arts and humanities scholars, for whom e-journals had little attraction[2]. In direct contrast, publishers have moved closer to the user as more and more information transactions are undertaken in their territorial space. Obtaining COUNTER[3] data in return is not really an acceptable or sufficient form of recompense for what is being lost.

About ten years ago, everybody was talking about 'disintermediation' – the Online Conference at Olympia was chock-full with papers. You don't hear that word used any more, which I suspect has put

information professionals off their guard, but it has happened big-time, and with greater consequences than were originally expected.

In an enormous, virtual, disintermediated information environment which is full of choice and where volatility is the normal condition, routine and real-time monitoring is clearly absolutely essential if we want to know what is going on.

Of course, librarians have been bleating on about users since time immemorial, but have not really made that much progress in getting closer to them. It is almost as if, by mentioning users, this assuages the guilt. How many libraries have a department dedicated to following the users' every move and relating that directly to academic outcomes and impacts for the benefit of senior academic managers? The answer is none that we know of, and this is seriously worrying given that, to succeed in the information business, we need to follow their every move. They are, after all, driving all the major changes in the digital information environment. The big challenge here for us all is in understanding and accommodating the concept of the digital information consumer. COUNTER, as at present constituted, is of very little help here as it tells us about hits (activity) but not about users.

Like it or not, libraries are a part of the consumer information marketplace, and they need to respond accordingly. Can you imagine another industry, or even the government, having so little information on its users? Look at supermarkets, Tesco for example: they find out whatever they can about the customer, and respond immediately. Tesco's success is largely attributed to their deep understanding of the customer obtained through the data collected from loyalty cards and the like. There is an example here for us all.

With massive and rapid change occurring on such a wide front, innovation is absolutely crucial; it is the key to survival. However, successful innovation depends on the ability to map and measure the impact of changes that occur with innovation. Did, for instance, change happen in the way expected? Digital roll-outs rarely impact in the ways expected. If you need an example of this, take mobile phones. It was originally thought that the market for them would be people remote from their bases, such as farmers and sales personnel. We now know that the really big market is children! Librarians need to innovate to just stay alive and stay relevant (and, to be fair, they have had a good record to date of doing so), but future success rests more than ever on robust and timely usage and user data, not hype or PowerPoint puff.

## But there is hope

That is the bad news; the good news is that it has never been easier to collect user information, as everything the user does in the virtual environment is automatically recorded, the complete opposite of what happened in the physical environment where an enormous effort had to be expended to obtain very selective information on an irregular basis. We have an opportunity to monitor use like never before, and in the process probably find out that we have made many an erroneous assumption about how users behaved. The book on the user can be put right, indeed, rewritten. Never easier to do and never more important, which makes it so puzzling that so little is done in regard to user evaluation.

As previously mentioned, COUNTER data can provide a start but it really just provides records of activity associated with an information source, not of information-seeking behaviour, which is what is really needed to plan information systems, judge information-seeking success, satisfaction and scholarly outcomes. The real need, however, is to move away from counting hits and downloads (activity) to counting users – through *deep* log analysis. And only when we have *user* data can we move on to those impacts and outcomes that will surely have to be produced to guarantee existing levels of funding, never mind increases. It is not generally known that logs themselves do provide us with user data[4]. There is name/type of institution for all to see; from a DNS look-up of the user's IP number we can glean type of organization and geographical location. Whether someone is a subscriber or not can also be discerned, and the person's subject field can be inferred by the subject of the material they are using. In some cases, sub-network labels contained in the logs can tell us the name of the user's department and whether they are searching from a student location (halls of residence, for instance). Some logs contain links to subscriber demographics, as those in the Blackwell Synergy study[5] did, and from this subject, gender, nationality and occupational status can be derived. It is also possible to link usage data to online questionnaire data for

the same person, as was achieved in the ScienceDirect study[6], but this is somewhat trickier to do. The potential information yield, though, is enormous.

CIBER, in its many projects and publications, has demonstrated how this can be done on laptops by a few individuals with the information-seeking, statistical and computing knowledge that can be found in many decent-sized libraries. The Research Information Network (RIN)-funded 'Evaluating the usage and impact of e-journals in the UK' study[7] (2008) and the JISC-funded 'UK national e-books observatory' study[8] (2008-2009) are currently taking deep log analyses to the academic community and sharing them with them. Hopefully, this will stir up an interest and librarians will start doing it for themselves.

## What does deep log analysis tell us about the information behaviour of the virtual scholar?

The descriptions that follow are based on the digital fingerprints (logs) of over 5 million scholars and over 20 million views, so we can say it is a robust evidence base of unprecedented size. Interestingly, it was publishers who first woke up to the benefits of the data; partly, a reflection of the fact mentioned earlier, that scholarly usage increasingly takes place in the publisher space.

Here, we pick out those findings that should interest librarians most. They concern: popularity, user diversity, 'bouncing', online reading, digital visibility, search engines, trust and outcomes. The data is variously taken from studies of ScienceDirect[9], Synergy[4], EmeraldInsight[10,11], OhioLINK[12,13] and Oxford Scholarship Online[14,15].

### *Popularity of scholarly information seeking*

Scholarly information appears to be in great demand. This is particularly the case with e-journals and we suspect this will prove to be true for e-books too. A few examples to illustrate this fact are given below:

- As far back as 2004, Synergy attracted more than 500,000 visits a month, recording around 5 million views as a result of the visits;
- In the case of OhioLINK, of their 6,000 or so e-journals, all bar five were used within a period of a month;
- Just one journal, *Nucleic Acids Research (NAR)*, recorded 17,150 downloads in a single month and usage increased by 150% in the space of two-and-a-half years; and it is not just subscribers from whom the demand is coming. A study of EmeraldInsight showed that two-thirds of visitors were non-subscribers, happy to look at abstracts and free materials. In the case of NAR, in fact, the majority of users were non-subscribers.

The high volume and the strong growth in use is really a result of improved access to the product, courtesy of big deals, search engines, broadband and wireless. Open access and institutional repositories will fuel this growth even further.

Demand is genuinely universal, with typically, for example, UK-based, financed and targeted services attracting more foreign than home use, as was found in studies of Intute[16], a scholarly information gateway and BL Learn, a site designed to help the younger scholar[17]. This of course raises some quite interesting and, possibly, embarrassing questions regarding national interest and benefits.

### *Diversity of scholarly information seeking*

The first thing we notice when we move away from counting hits to counting users is the difference in individual information-seeking behaviour. There are very significant differences between various types of user, especially in regard to their subject field. Differences have also been found in regard to academic status, geographical location, gender, type of organization worked for, type of university and attitudes towards scholarly communication. Be wary of the stereotypical impressions conveyed by hit counts; one size does not fit all. Here are some examples from the ScienceDirect study[9] that help to prove the point:

- *Age of material* Users from economics (71%), engineering (71%), the social sciences (69%) and computer science (70%) made the most use of current (one-year old) material; those from material science (51%) and mathematics (52%) the least.

- *Number of journals consulted* Users from material science (39%) and mathematics (38%) were the most likely to view two or more journals in a session. Those from medicine (69%) and computer science (69%) were more likely to view just one.
- *Return visits* Computer science (80%) and physics (79%) recorded the highest percentages of repeat visits, and engineering the least (46%)
- *Age* 14% of those aged 36-45 undertook an abstract-only search; it was double that for those aged over 56.
- *Gender* Men were much more likely to undertake a session in which an article was only viewed in PDF form; the percentages were 37% (men) and 22% (women).
- *Geographical location* Eastern Europeans (47%) and Australasians (82%) recorded a relatively high percentage of searches resulting in zero returns. North Americans appeared the most 'successful' searchers: 74% of their searches resulted in one or more matches.

### *Horizontal information seeking ('bouncing')*

Probably the most notable and interesting characteristic of information seeking in the virtual environment is that, aided and abetted by search engines, much of it tends to be horizontal and fleeting in nature[18]. Thus:

- A half to two-thirds of website visitors typically view no more than a page or two during a visit and then leave;
- Many do not return: thus in the space of a year it might be expected that half would not come back. For a good number, user loyalty is at a premium, partly because of a dependency on search engine-searching, and choices here are always being refreshed;
- Users, plainly appreciating the huge choices on offer in the virtual environment, search a variety of sites to find what they want.

At *best*, we can conclude that this represents the adoption of a powerful checking/comparing, dipping sort of behaviour, which is a result of search engines, a shortage of time, gateways and huge digital choice. At *worst*, it represents a massive failure at the terminal, especially, perhaps, in regard to particularly vulnerable groups, like the young and those only recently introduced to the delights of database searching. The following section suggests that it might indeed be the latter conclusion that must be drawn.

### *Little 'reading' appears to be occurring online*

Bouncing raises questions about scholarly outcomes, whether they are what you might expect from all the activity that is patently going on. So, too, does another information-seeking characteristic, time spent online viewing a page or on a visit. People spend seconds viewing a page and many sessions last no longer than a minute or two. Online attention spans seem low: a) people spend more time reading shorter articles online than long ones; b) as the length of a paper increases, the greater the likelihood that it will be viewed as an abstract and the less that it will be viewed in full text. You might say that this is not really surprising as the web is as much a media form as it is an information resource; possibly, people go online to avoid reading. However, there is no evidence to suggest that all, or most, of what people download or print off is read at a later date; indeed, evidence, albeit anecdotal, suggests that much is not. Perhaps, what people are hoping from downloading is some form of digital osmosis. If this is indeed the case, then what of the value of that most prized metric, full-text downloads?

### *Digital visibility has an enormous impact on information seeking*

In a crowded information space where only a minute amount of content can be viewed at any one time, this is not surprising, of course. The way that improved access and increased visibility leads to increased exposure is no clearer than in the case of the use of journal back-issues. Thus, far from use being dominated by the current issue as we have been taught to believe, in fact, the ScienceDirect study showed that downloads of material older than five years accounted for very healthy volumes of use in the case of two scientific fields, materials science (59%) and physiology (64%).

Perhaps the most staggering finding in this regard was what happened when two Emerald journals were given high visibility by making their contents freely available for a week (Figure 1). Use increased a hundredfold, only to fall back to previous low levels of use when titles returned to subscriber control. What was really interesting was that, no matter what the journals were (new titles were offered each week), the effect was the same.

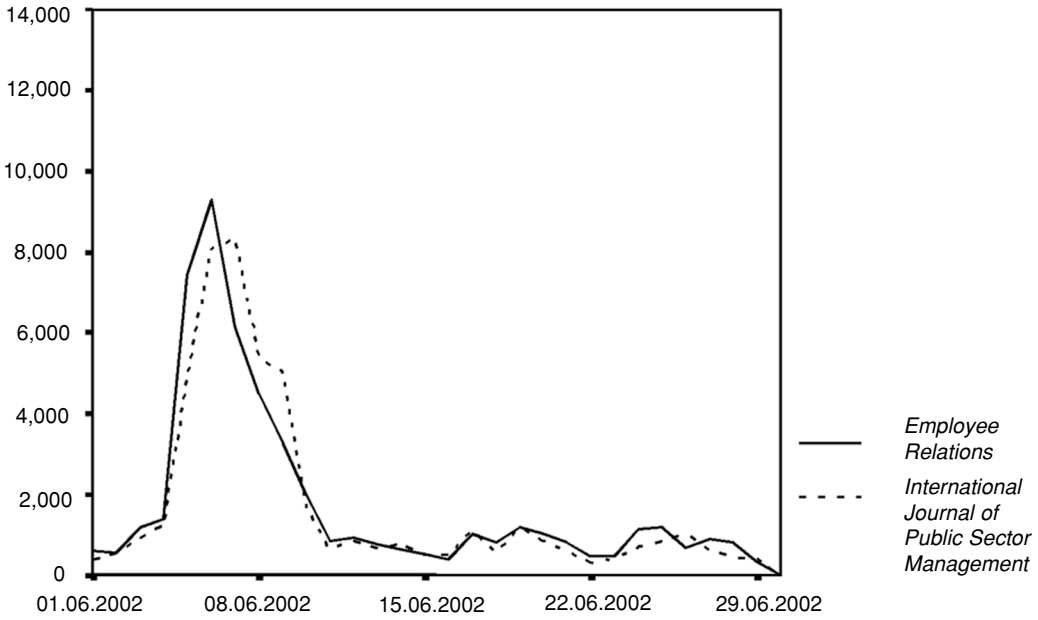


*Figure 1. Impact of two Emerald journals being offered free for a week*
*Source: Nicholas et al, 2003*

### *Search engines and information seeking*

We have already chronicled the great impact that ubiquitous search engine-searching is having on information-seeking behaviour, but there is much more. The information-seeking behaviour of search engine users is quite different and certain groups are more likely to use search engines. Thus it has been found that people using search engines were: a) far more likely to conduct a search that included a view to an older article; b) more likely to view more subject areas, more journal titles, more articles and abstracts too; c) more likely to be 'bouncers'; and d) more likely to be undergraduates.

Figure 2 shows the impact of opening up a scholarly web site to search engines, and what an impact it had. *Nucleic Acids Research* can hardly be described as a 'popular' journal, a journal that might attract a casual or passing user. You would think that its audience would be relatively circumscribed. However, the figure shows that as a result of opening the site to search engines in 2003, use grew by 150%. By comparison, when the journal went open access (denoted by the dotted line), this only had a marginal impact on use; it grew by less than 10%[19].

### *Trust and information seeking*

Librarians naturally set much store by authority and trust, believing that scholars need/want it, and that they, and the systems they provide on behalf of them, provide it – see for instance the marketing associated with gateway sites, like Intute, which portrays itself as a safe haven from the dangers of search engine-searching. However, in cyberspace, authority is not easily ascribed because there are so many players involved in trust and authority judgements and therefore judgements are not so easily made[20]. And here lies a very big problem for libraries. Take the example of a virtual researcher searching from their office in a university. They have conducted a Google search, as increasingly large numbers of them do, to find the Synergy database. On connection, a cookie identifies them and provides them with full-text access. Now,

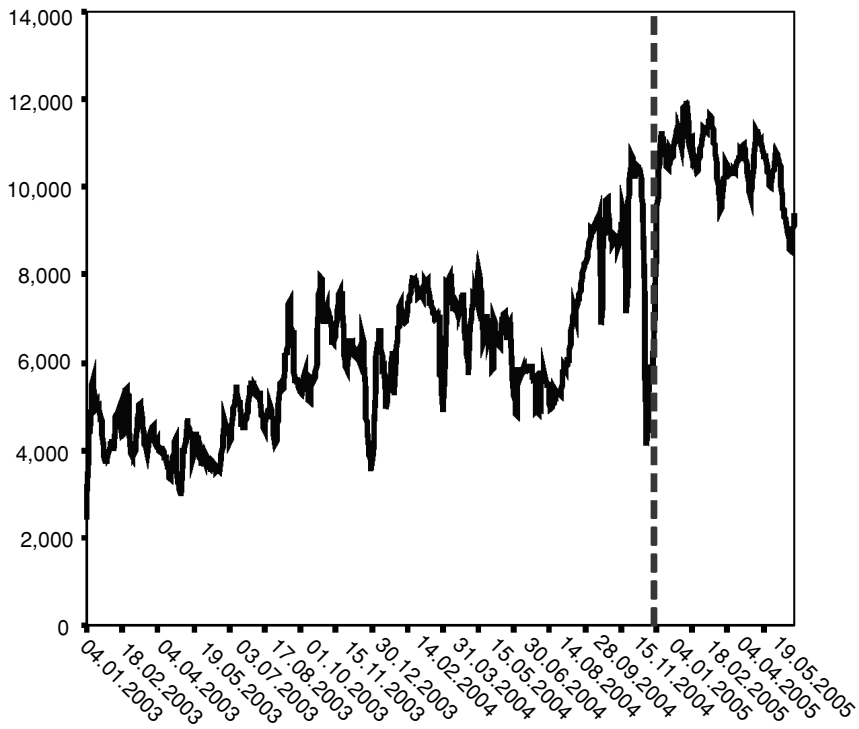


*Figure 2. Nucleic Acids Research - daily article views from January 2003 to May 2005*

that researcher used a Microsoft browser, then Google, then Synergy, then arrived at the *Journal of Computer-Mediated Communication* and, on inspection, alighted on an article by Smith from City University. Throw into the pot that they might or might not have known that: a) the library had paid the subscription, so providing full-text access; b) that Synergy was produced by Wiley-Blackwell; and c) that the journal was published on behalf of the International Communication Association. In these circumstances, where does the authority lie and how is the user meant to make sense of all this in a visit that is characterized by its fleeting nature?

The evidence in fact tells us that many users are promiscuous, assess authority and determine trust very quickly (in a matter of seconds) by cross-checking (bouncing tells us this) and, if they do notice kite marks, assurance labels and the like, they only do so very fleetingly. Furthermore, the younger they are, the less trusting (and more promiscuous) they are, and the less likely to recognize traditional brand names. Indeed, for a good number of them, Google is a far stronger virtual brand than, say, Synergy and if they thought that Synergy was in fact part of the Google family they would probably trust it more than if they knew it came from Wiley-Blackwell and that it was a publisher. Authority (and relevance) has to be won (and checked) and at present, unfortunately, we have little evidence regarding the authority a library possesses in cyberspace.

### *Outcomes*

In an information world where journals dominate the agenda, it is surely incumbent on us to assess how the unprecedented and (increasingly) costly access to e-journals by the academic community has influenced their information-seeking behaviour, whether it has led to an improvement in the research process and scholarly outcomes. We all make the assumption – librarians in particular – that increased access is hugely beneficial in some way or another, but nobody has yet sought to establish this in an evidence-based way. Thanks to COUNTER data, we all know that there is a lot of activity associated with many scholarly e-journal databases/publisher platforms, which lulls everyone into a general sense of well-being, but, unfortunately, COUNTER logs and the like cannot ascribe this activity to: a) information-seeking behaviour; b) communities of users, such as physicists, economists, research-intensive departments, etc; c) individual articles/papers. Nor can we distinguish between student and staff use. Without

this information we cannot know in what way the tremendous strides (and costs) in providing access to e-journals have benefited the research community. Deep log analysis does enable us to do all this, and in the RIN-funded project 'Evaluating the usage and impact of e-journals in the UK', 2008, this is precisely what CIBER will be doing.

## Conclusions

Fantastic access, an abundance of scholarly information, huge digital choice and a common/multi-function retrieval platform are changing everything; scholars are now consumers in every sense of the word and they are exercising their new found powers and wings, and this leads to increasing levels of information promiscuity. This is transforming their relationship with all information providers, and weakening that with the library. Without the necessary user or market data and general day-to-day contact with users, the response of libraries has tended to be a technological, rather than a user one, and this is going to bring big problems unless the situation is quickly rectified. How many commentators have said this in the past and yet we are still saying it today? However, the stakes are much higher now and the risks of complete meltdown never more real. The advice has to be, put technological innovation on the back-burner and put user evaluation on the front-burner. A good start can be made by harvesting the information-seeking data contained in the logs that are so abundantly being produced by digital information systems, remotely and without massive effort. Publisher help will be needed but, of course, disputes arising over open access publishing have complicated matters.

The Doomsday scenario surely has to be: a) the digital transition sees users fly into the hands of others (publishers); b) e-books fast-track this process; c) as a result, libraries obtain even less credit for what they do; d) at the same time, the impacts and outcomes of online access are questioned by Faculty.

## References

1. CIBER projects:
   http://www.ucl.ac.uk/slais/research/ciber/projects/ (Accessed 28 January 2008)
2. Nicholas, D, Huntington, P, Rowlands, I, E-books: how are users responding? *Update*, 2007, 6(11), 29–31.
3. Project COUNTER:
   http://www.projectcounter.org/ (Accessed 28 January 2008)
4. Nicholas, D, Huntington, P, Watkinson, A, Scholarly journal usage: the results of deep log analysis, *Journal of Documentation*, 2005, 61(2), 246–280.
5. The Virtual Scholar research programme – use and impact of digital libraries in academe. Funded by Blackwell/Emerald/Elsevier; 2003–2004.
6. Authors as users: a deep log analysis linking demographic and attitudinal data obtained from Elsevier authors with their usage of ScienceDirect. Funded by Elsevier; 2005–2006.
7. Evaluating the usage and impact of e-journals in the UK:
   http://www.rin.ac.uk/use-ejournals (Accessed 28 January 2008)
8. UK national e-books observatory:
   http://www.jiscebooksproject.org/ (Accessed 28 January 2008)
9. Nicholas, D, Huntington, P, Jamali, H R, User diversity: as demonstrated by deep log analysis, *Electronic Library*, 2008, 26(1), [In Press].
10. The Virtual Scholar research programme – use and impact of digital libraries in academe. Funded by Blackwell/Emerald/Elsevier; 2003–2004.
11. Nicholas, D, Huntington, P, Watkinson, A, Digital journals, big deals and online searching behaviour: a pilot study. *Aslib Proceedings*, 2003, 55(1/2), 84–109.

12. Maximizing Library Investments in Digital Collections Through Better Data Gathering and Analysis (MaxData). Funded by US Institute of Museum and Library Services; 2005–2007.
13. Nicholas, D, Huntington, P, Jamali, H R and Tenopir, C, Finding information in (very large) digital libraries: a deep log approach to determining differences in use according to method of access. *Journal of Academic Librarianship*, 2006, 32 (2), 119–126.
14. SuperBook. Funded by Emerald and Wiley Publishers; 2006–2007.
15. Nicholas, D, Huntington, P, Rowlands, I, Dobrowolski, T, Jamali, H, Superbook: an action research project. *Online Information 2007 Proceedings*, 2007, 50–57.
16. Intute:
http://www.intute.ac.uk/ (Accessed 28 January 2008)
17. The Behaviour of the Researcher of the Future (Google Generation). Funded by the British Library and JISC, 2007.
18. Nicholas, D, Huntington, P, Jamali, H R, Dobrowolski, T, Characterising and evaluating information seeking behaviour in a digital environment: spotlight on the 'bouncer', *Information Processing & Management*, 2007, 43, 1085–1102.
19. Nicholas, D, Huntington, P, Jamali, H R, Open access in context: a user study, *Journal of Documentation*, 2007, 63(6), 853–878.
20. Nicholas, D, Huntington, P, The virtual scholar, Online Information, 2006, 19–2.

© *David Nicholas*

■ **Professor David Nicholas**
**Director, School of Library, Archive and Information Studies and**
**UCL Centre for Publishing**
**University College London**
**Henry Morely Building**
**Gower Street**
**London WC1E 6BT, UK**
**Tel: +44 (0)20 7679 2477**
**Fax: +44 (0)20 7383 0557**
**E-mail: david.nicholas@ucl.ac.uk**
**www.ucl.ac.uk/slais/david-nicholas/**

## Biographical note

David is Director of the School of Library, Archive and Information Studies (SLAIS) at UCL, Director of the UCL Centre for Publishing, and Director of CIBER (Centre for Information Behaviour and the Evaluation of Research). His research interests largely concern the virtual scholar and he is currently engaged in research investigations of e-journals (usage of open access OUP journals; OhioLINK, RIN study on e-journal usage and outcomes), e-books (JISC-funded national e-book observatory) and young scholars (GoogleGeneration). Previously, David was Head of the Department of Information Science, City University.

To view more chapters from *The E-Resources Management Handbook*, published by UKSG, click here:

**http://www.uksg.org/serials/handbook.asp**