Information: Needs for the Future

Krys Bottrill

FRAME, Russell & Burch House, 96–98 North Sherwood Street, Nottingham, NG1 4EE, UK

Summary — The four central questions surrounding the use of information are: where to find it; how to find it; how to present it; and how to maintain information availability and information literacy. It is usually assumed that the main source of information for most scientists is the peer-reviewed journal literature. Traditional journal publishing is beset with a number of problems. Although electronic publishing might possibly solve some of these, it in turn introduces new problems. Further problems arise with respect to secondary sources which, in some cases, are being supplemented by electronic archives of full-text documents. One fundamental question that arises when considering any large collection of documents or of records about documents is whether or not to index them, and how to index them. The pros and cons of free-text searching versus the use of controlled vocabularies are discussed, as is the importance of harmonising the Three Rs-related terminology of existing and proposed thesauri. However, there is a further problem that documents pertinent to the Three Rs are not always indexed from this point of view. Authors need to be made aware that, if the information is not provided in the abstract, there is no easy way to identify and retrieve this document from a database. Small specialised databases on the Three Rs in relation to specific subject areas could provide a further solution, especially if they provided references to conference proceedings and book chapters, which are not usually found in the large bibliographical databases. The provision of training in the use of information resources, and the establishment and maintenance of these resources, require investment of money and professional skills. Finally, the future of Three Rs information depends on a recognition that this in an important topic which deserves more than lip service.

Key words: *databases, information, journals, publishing, thesauri*.

Introduction

My title, as originally given, was *Databases — Needs for the Future*; however, I have taken the liberty of changing this to *Information — Needs for the Future*, because the existence or non-existence of databases, and the types of content presented, form only a part, albeit a vital part, of the problematical area of communicating and finding information that is required for any specific purpose, including the purpose that is of greatest concern to this audience, namely, introduction of the Three Rs into all areas of biomedical research and testing.

I intend to look at four central questions: where to find information; how to find information; how to present information; and how to maintain information availability and information literacy. I will be looking at how these questions operate generally and in the scope of information on the Three Rs, and will use them to form a wish list for future developments.

Where to Find Information

The first question faced by someone needing to obtain information is where to look for it. It is usually assumed that the main source of information for most scientists is the peer-reviewed journal literature. The greatest problem associated with journal publication is that there can be a considerable delay between a scientific study being conducted and its publication in a journal.

A Web site which collects current information about journals from scientists publishing in the field of computer science found that the time lag between submission of a paper to a journal and its eventual appearance could vary from 9 to 26 months (1). A 2001 study from Princeton on publication delay in the same field, estimated by using the average age of the most recent citation in articles at the time of publication, gave a range from as little as 0.5 months for some workshop proceedings, to as much as 39 months for the worst-performing journal, and 10 months for the best performers (2). Thus, even when not allowing for the delay that occurs before the contents of a scientific journal are indexed into a database, "current" scientific literature is in many cases not particularly current.

Moreover, it is debatable whether direct consultation of the published literature does play a significant role in the information-gathering activity of scientists. An interesting study among academic psychologists, conducted in the early 1960s by Garvey & Belver Griffith (3), found that the appearance of research results in a journal accounted for only a small percentage of the information communication among scientists. It typically involved a time-lag of at least 18 months between the start of a project and the first publication of results. Over 50% of papers in a core journal would be read by less than 1% of a random sample of psychologists,

and even the most-consulted would be read by no more than 7%. However, 40% of scientists alerted others to their work by distributing preprints, and 62% by distributing reprints, although the average author distributed only ten preprints of each paper published. This is one reflection of the functioning of the Invisible College, the network of informal and elitist scientific communication first described by the late and sorely missed Derek de Solla Price (4). It is also through the Invisible College that select groups of scientists communicate their opinions about current theories, practice of methodology, and other topics which do not readily find their way into the formal published literature. However, the Invisible College is an exclusive structure, which is not readily accessible to all scientists, especially those outside the main centres of research.

Electronic publication is one strategy that might speed up the publication process, although the refereeing process would continue to be a bottleneck. Electronic publication might also open up the Invisible College to wider participation. A paper can be published as soon as it has been accepted by the referees, rather than having to wait until there is space for it in the publication schedule (5). For example, the *Journal of Biological Chemistry* publishes papers online on the day they are accepted, and before they have been copy-edited for the hard-copy edition. On average, this means that the paper is accessible about eight weeks before it appears in the paper edition (information from the journal Web site at http://www.jbc.org/pips/index.dtl). Likewise, it is also technically possible for authors to make their preprints more easily accessible by placing them on the journal's Web site even before the paper has been formally refereed. In addition, the electronic format opens the way for a rapid debate between authors and readers about the contents of a paper; something that, because of the time-lag of publication, is not truly feasible in the letters pages of a conventional hard-copy journal.

Another potential advantage of electronic publishing is the possibility to archive all the raw data and other associated information together with the paper (5). This is of particular interest to the Three Rs, because it would permit the archiving of detailed methodological information and protocols, including details of how refinement, reduction and/or replacement have been implemented in the study. This could have a great impact on encouraging the use of alternative approaches.

However, electronic publishing brings its own problems. One great concern is that, paradoxically, there is a danger that the accessibility of information will be decreased if scholarly electronic publishing is conducted as a commercial venture. A journal in hard copy, once purchased, remains the permanent property of the library or individual buyer, and is available for as long as it is considered valuable enough to justify its shelf space. In contrast, access to the archives of an electronic journal is limited to current subscribers, so if a library decides not to continue with a subscription, all the past information from that journal, which has already been purchased, will be lost to its users. Similarly, information will be lost if a publisher ceases to trade and shuts down its Web site.

Many are critical of the current system of scientific publishing, be it on paper or electronic. To quote one such critic, it is a system in which "research is funded largely by the public, scientists turn over their intellectual product freely to a commercial third party, and are then forced to buy it back at a premium price" and it is "failing fast" (6). A number of initiatives are attempting to provide a solution to the problem.

The National Center for Biotechnology Information (NCBI) at the US National Library of Medicine (NLM) has established PubMed Central (http://www.pubmedcentral.nih.gov/), a digital archive of life-sciences journal literature, with free and unrestricted access. Unfortunately, presumably due to pressure from publishers, PubMed Central no longer requires that the material be deposited in its own archive, but will allow publishers to provide links to the full text on their own sites, thus negating its objective of providing a secure archive.

The Public Library of Science project has so far collected the pledges of nearly 30,000 scientists in 177 countries that they will publish in, edit or review for, and personally subscribe to, only those scholarly and scientific journals that have agreed to grant unrestricted free distribution rights to any and all original research they have published, through PubMed Central and similar resources, within six months of the initial publication date. The organisers have also stated their intention to establish a non-profit scientific publisher operated by scientists. They are starting to form an editorial board and to raise money to cover the initial operating costs (http://www.publiclibraryofscience.org).

BioMed Central (http://www.biomedcentral.com) is an already-existing independent, electronic publisher "committed to providing immediate free access to peer-reviewed biomedical research". It consists of more than 50 online, peer-reviewed journals that publish papers as soon as they are accepted, simultaneously entering them into the PubMed archive. (It is interesting to note that medical articles undergo an open peer review, with the signed reports of the reviewers being posted onto the Web site together with the article, while biological papers are still reviewed anonymously.) The policy of Biomednet is to enable free *access* to information; therefore, publication charges will be imposed to cover the costs of processing and storing the articles, although these will be waived in cases of need. This is not so great a drawback, since many commercial journals already impose page charges on authors, as well as making readers pay for access.

Thus, it seems as if there might be a move back to a situation where scientists take control of scientific publication and return it to a non-profit basis, which, in my opinion, would be a most welcome development. However, the establishment of vast electronic archives raises another potential problem, that of information retrieval, which is the next question I wish to consider.

How to Find Information

Traditionally, when it was necessary to go beyond personal knowledge plus informal sources of information, such as colleagues, the scientist would turn to the secondary literature. Secondary sources are those which present information about the primary literature, i.e. about the contents of journals, conference proceedings and books. Secondary sources originally took the form of hardcopy collections of abstracts, indexed by subject matter and by author. Today, these have for the most part been replaced by computerised literature databases available through the Internet or on CD-ROM. In addition, it is also possible to search directly in an archive containing the full texts of deposited documents.

One fundamental question that arises when considering any large collection of documents or records about documents is whether or not to index them, and how to index them. This is a debate which has been going on for years between information technologists and information scientists. The options are to have no indexing at all, to use automated indexing systems, to employ people to do the work manually, or to combine these approaches.

I have heard a number of IT people express the opinion that, as computers become faster and more powerful, they increasingly make indexing redundant. They argue that, since it now only takes a few seconds to conduct a free text search across a massive collection of full-text documents to find the ones that include the search term, there is no need to spend time and money on indexing. Free-text searching can certainly be useful in some cases. Its main advantage is flexibility. Any word can be used as a search term. This means that, in rapidly advancing subjects, free-text searching will identify articles using new terminology before indexers have caught up with developments and included these terms in their thesauri. Free-text searching is also appropriate for very specific name searches, such as the name of an assay system. However, free-text searching also has many pitfalls. It does not permit the search for a term to be limited by concept. The best example of this can be seen when one searches the Web, which is the largest unindexed archive of information in existence. All of us have experienced the frustration of putting words into a search engine and then having to spend much time sifting out a few valuable finds from a mass of irrelevant information. Moreover, unless one can be confident about putting in all possible ways in which an idea can be expressed, there is no guarantee that all the information will be found through a free-text search. An early study of searches for legal information, where it is frequently important to search exhaustively, showed that lawyers using a database containing 40,000 documents considered that it was acceptable to retrieve 75% of all relevant documents, and thought they succeeded in this by using free-text searching. However, an analysis of their results showed that they were actually retrieving only 20% of all relevant documents (7).

Information specialists prefer to use free-text searching as an adjunct to searches using the indexing system of a database. There is another ongoing debate about the relative merits of automatic and manual indexing. Automatic indexing has the merits of never missing a term, and it can map synonyms to the preferred term in a way transparent to the user. Thus, someone searching with the phrase “dairy products” can be presented with documents containing the words “milk, “cheese”, “cream”, “butter”, etc., even if they do not contain the phrase itself. However, automatic indexing can only follow rules, and so would also present a document as relevant which might have nothing to do with dairy products, but contained a reference to the “milk of human kindness”!

Obviously, the question is one of economics: should the monetary cost come at the level of data input, with a trained indexer paid to allocate terms to describe the concepts contained in each document; or should it come at the level of data retrieval, in the cost of time spent filtering out a mass of irrelevant material?

I would like to see all online Three Rs information resources sufficiently well-financed to ensure that they are properly indexed by competent human beings, rather than by machines. This would increase the ease with which users could find the information that they required, and therefore would encourage the use of these resources, to the benefit of scientists and animals alike.

If we accept the need for manual indexing, this in turn raises the question of how the indexing is done. The first and most vital aspect is the controlled vocabulary that is used for indexing. One of the major problems in finding Three Rs information in the large, general literature databases is the lack of Three Rs-related terms in the thesauri used to index them. Some progress has recently been made in this respect. The MeSH thesauri used to index Medline now includes the term “animal testing alternatives”, which can be used with a number of qualifiers to index documents about the Three Rs more precisely than was previously possible. Meanwhile, The US Department of Agriculture’s Animal Welfare Information Center (AWIC) has produced a thesaurus for animal use alternatives of

over 200 terms, to improve indexing of these topics in the Agricola database (http://www.nal.usda.gov/awic/alternatives/altfact.htm). ECVAM's Scientific Information Service (SIS) is also involved in the production of a Three Rs thesaurus. I would like to see some attempts at harmonisation between these and other similar initiatives in the future, so that all indexing of the Three Rs can at least proceed from a common starting-point.

Even with the best thesaurus in the world, documents pertinent to the Three Rs will not be found if they are not indexed as such. Most databases limit the number of index terms that can be assigned to any one document. In addition, indexers are usually not permitted to index anything that is not explicitly connected with the stated objective of the document. Therefore, even if the authors mentioned in the methods section that they selected the methods used as an alternative to using an animal system, the document might not be indexed with any Three Rs terms, if the main topic of the article were something other than the presentation of an alternative method. If the information was not provided in the abstract, there would be no easy way to identify and retrieve this document from a database.

For this reason, I think it is useful to have databases specifically about the Three Rs. However, I do not think it is possible to construct one universal database for this purpose, since this, in effect, would mean reindexing the whole of the biomedical literature. In my opinion, the best way forward is to establish smaller databases restricted to specific subjects or disciplines. Examples of such databases already available include the Altweb database on anaesthesia and analgesia, the Norina database on alternatives in education, and the ECVAM SIS databases on alternatives in toxicology. It would be useful for any subject-specific databases to obtain input from scientists who are active in the field and familiar with alternative approaches. In most cases, the starting point will be the establishment of a bibliographical database of published literature. Extra information, such as method protocols, potential sources of funding, and contact details for people working in that area and interested in developing alternative approaches, would, however, add to the value of the resource. It would be extremely useful for these small, specialised databases to include references to conference proceedings and book chapters, as well as to individual papers, since the former types of information are not usually included in the large, general, biomedical databases.

It is of little use to have specialised resources if nobody knows about them. Therefore, it is important to have referral points from which to survey and access the information that is available. The Altweb Web site is well on the way to becoming such a point of reference. Apart from archiving a lot of documents on its own site, Altweb also links to numerous Three Rs-related Web sites around the world.

In a new development, it is now possible to use the Altweb search engine to perform a simultaneous search on all these sites. Altweb still requires a lot of refinement of its functions, but it is already a very valuable resource. In my opinion, one of the most pressing needs for the future of Three Rs information is to guarantee long-term financing for Altweb, so that its existence and further development are secured. It would also be very useful to establish mirror sites of Altweb outside the USA, both to reduce bandwidth use and to provide a backup service. A mirror site could also offer region-specific information on the Three Rs alongside the main Altweb content.

How to Present Information

I have already referred to the problem that, with respect to mainstream databases, a document will most probably not be indexed with Three Rs-related indexing terms if these do not describe its primary objectives. Therefore, the only way to ensure that it will be retrieved from a bibliographical database is for authors or journal publishers to ensure that in such cases, a reference to reduction, refinement or replacement is made in the abstract. It is not enough to indicate these in the key words added by the journal to the paper. Database indexers are often instructed to ignore journal-assigned or author-assigned key words when indexing a document. However, there is the difficulty that the phrase "Three Rs", and the words "reduction", "refinement" and "replacement" can be used in multiple contexts, and therefore are not always very useful as search terms, even when linked with subject-specific terms. It would be useful if we could agree on terms that are more likely to be unique. In my view, the terms used in the AWIC thesaurus, "animal use reduction", "animal use refinement" and "animal use replacement" are phrases that could usefully be used for this purpose.

Likewise, it would be useful if all those involved in placing Three Rs-related information onto the Web agreed to include relevant selections from an agreed terminology in their page content. Preferably, such terms should also be included in the metadata associated with the pages. In this context, I think it might be useful for information providers to discuss whether to introduce the Dublin Core Metadata Elements (8) into their Web pages.

How to Maintain Information Availability and Information Literacy

Finally, the future of Three Rs information depends on a recognition that this is an important topic which deserves more than lip service.

Scientists need training in how to search for information on alternatives to animals. A study of information professionals in corporate, scientific, technical and academic fields, commissioned by *Information World Review* and *Dialog*, identified "information illiteracy" of database end-users as a major problem which harms the effectiveness of their organisations (9). This was particularly marked in the academic sector, where 65% of respondents expressed this concern. This overall deficiency in search skills will inevitably be reflected in searches related to the Three Rs. There is a general consensus in discussions between librarians, such as on the lis-link mailing list, that information departments, in particular academic libraries, have insufficient resources to offer adequate training in search skills to library users. Moreover, in relation to the mandatory courses currently undergone by personal and project licence applicants in the UK, the timetable does not permit extensive treatment of the Three Rs, let alone tuition in how to search for information on this topic. For all these reasons, the provision of lectures, workshops and interactive tutorials on search techniques and information resources specific to the Three Rs, would be a valuable contribution to promoting the use of alternatives by scientists.

Both the provision of educational resources and the establishment and maintenance of databases require investments of money and professional skills. There is little point in setting up a database if the money to maintain and update it on a long-term basis is lacking. This is a perennial problem, which can only be solved through a concerted effort by the major funders of Three Rs research. Ultimately, it is in the interests of the funders to ensure the existence of such resources, because there is little point in funding research without at the same time ensuring that the results of the research are made easily available to the scientific community and to regulators.

Hau & Carver (10) called refinement the Cinderella of the Three Rs. However, in my view, it is information that is the true Cinderella. The topic of information has not been deemed worthy of its own section in the three World Congresses on alternatives held so far, nor in the forthcoming Fourth Congress. Instead, it has to share a timetable with the topic of alternatives in education. Likewise, it has never been the subject of a plenary lecture at any of the congresses. At the very least, I would like to see a satellite day added to the congresses, which would focus on information. This could include educational sessions, aimed at information users, on how best to conduct searches and what resources to use for searches. More importantly, however, it would provide the opportunity for information providers and information specialists to explore how best to make use of what is already available and to define priorities for future initiatives. It is time to stop paying lip service to the subject of information, and to recognise its central importance in the development, promotion and acceptance of Three Rs alternatives.

References

1. Hutter, M. (2002). *Review and Publication Time of Computer Science Journals*. Web site of Istituto dalle Molle di Studi sull Intelligenzia Artificiale http://www.idsia.ch/~marcus/journals.htm#going (Accessed 23.4.02).
2. NEC Research Institute (2001). *Estimated Publication Delay (Lower is Better) — Research Index* November 2001. Web site http://citeseer.nj.nec.com/pubdelay.html (Accessed 23.4.02)
3. Garvey, W.D. & Belver Griffith, B. (1964). Scientific Information Exchange in Psychology. *Science* **25**, 1655–1658.
4. de Solla Price, D.J. (1986). *Little Science, Big Science . . . and beyond*. 301 pp. New York, NY, USA: Columbia University Press.
5. Chan, L.K.W. (1996). Exciting potential of scholarly electronic journals. *CAUT Bulletin* **43**, 9.
6. Markovitz, B. (2000). What's wrong with how we judge science? In *Proceedings of the Freedom of Information Conference: the Impact of Open Access on Biomedical Research*, 6–7 July 2000, New York Academy of Medicine. Web site http://www.biomedcentral.com/info/conference.asp (conference home page); http://www.biomedcentral.com/info/markovitz-ed.asp (this article) (Accessed 13.5.02).
7. Blair, D. & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the Association for Computing Machinery*, 289.
8. Dublin Core Metadata Initiative (1999). *Dublin Core Metadata Element Set, Version 1.1: Reference Description.* Web site http://dublincore.org/documents/dces/ (Accessed 22.5.02).
9. Anon. (2001). The Pulse. *Information World Review*, September 2001 (issue 172), 1,3,18.
10. Hau, J. & Carver, J.F.A. (1994). Refinement in laboratory animal science: is it a Cinderella subject, and is there conflict and imbalance within the 3Rs? *Scandinavian Journal of Laboratory Animal Science* **21**, 161–167.