# Academic libraries and the challenge of abundance: the impact of the explosion of retrievable information on universities

**John MacColl**[1]
Online Computer Library Centre Research, St Andrews University Library, U.K.

## Introduction

Two months ago, Daniel Greenstein, Vice Provost for Academic Planning and Programs in the University of California, told a meeting of library directors in New York that:

> "*the university library of the future will be sparsely staffed, highly decentralized, and have a physical plant consisting of little more than special collections and study areas*" [1]

What he appeared to be suggesting was that, in a maturing digital world, information users no longer require librarians to mediate between the tools of information retrieval and their search needs. His comments caused a number of librarians to retaliate with evidence that their efforts in information literacy are still highly appreciated. However, this role is increasingly a contested one; particularly in view of the pitch that user empowerment has now reached, together with the new requirement for libraries to manage abundance and the priorities that arise from that.

## The scarcity model

Academic libraries have experienced significant change in their role in relation to information retrieval since the mid-1980s, when the computer was just beginning to make a major impact on how libraries provided information. What do we mean by 'explosion'? We are talking about changes that have affected universities over the past 40–50 years. The backdrop is massification in the higher education system, the rise of 'big science' and the transformation in the technologies of communication and publication wrought by, first, the computer, then the network, then the web.

The academic library of the mid-1980s in the U.K. would typically contain only one or two personal computers, or what were then called microcomputers,

[1]Email: maccollj@oclc.org

 

which were used by the qualified subject librarians to 'run computer searches' on behalf of users. This was then the cutting edge of the retrievable information explosion. What had happened in the previous few years was that the bibliographic data publications, which were the main sources of aggregated information on scientific disciplines, had been turned into computer databases. Initially used to generate printed output, we had by the mid-80s reached a point where convenience was best served by querying these databases from a computer over a network. We had moved, in effect, into the era of the 'online library'. It was bolted on to the print library. However, we were still performing the essential task of all libraries at all times, managing scarcity.

This had an impact on user behaviour and on library service. There was also an impact on the fundamental model of library operation. Users could still, if they chose, do their own literature searching using the printed databases. Humanists would want to search for books, using the major aggregated sources of the British Library Catalogue, the National Union Catalogue, or the British National Bibliography. Scientists were more interested in the journal literature. Chemists would use Chemical Abstracts. There were several other abstracting and indexing journals, typically beginning to fill many book stacks in large university libraries. Then there was an exceptionally interesting publication, the Science Citation Index, from the ISI (Institute of Scientific Information).

## From convenience to the newly possible

The Science Citation Index was interesting because it represented a paradigm shift in the delivery of aggregated metadata. Previously, databases had been computerized versions of printed publications. The computerization had brought the benefit of collapsing the chronological units, the monthly or quarterly printed journals, aggregating them into annuals and sometimes larger aggregations, so that the computer database could allow searches over multiple years. This was a leap forward in convenience. But the Science Citation Index did something new that had no parallel in the print-based tools. It analysed the journal articles that it indexed, parsing out the citations they contained, and allowing those citations to generate their own indexes. In effect, it data-mined the corpus and produced a new tool. It thus became newly possible not only to see the list of citations that an article contained within its metadata, but also to search the entire database by these citations. This was important because it revealed scientific impact for the first time. It became possible to start with a particular paper, or author, and see what the impact had been on the subsequent literature, demonstrated by the number of times a paper had been cited. Not only that, it became possible to read a topic by impact, going directly to the citing articles. What was revealed was a chain of influence of scientists upon each other, and this chain could be navigated, either forward or backward, by the person making the query. The effort of creating a tool like this for the printed literature would have been overwhelming, especially as the output of science was growing dramatically every year. Information science had reached a new point, in which one of its major tools was now *dependent upon* computing power for its very existence. It had no pre-computer analogy.

The Science Citation Index appeared as a computer database originally under the name Web of Science (now part of a bundled product from Thomson Reuters called Web of Knowledge). At this point, the early tools of bibliographic enquiry, which were simply electronic versions of printed publications, began to give way to new tools that exploited computing power in interrogating textual datasets. This shift, together with the flood of digital full-text information, moved the 'online' or 'electronic' library to one in which information in electronic form became predominant in some areas (e.g. journals in science, technology and medicine), search software became 'born digital', in the way of Web of Science, and library services eventually began to conform to the dominant idiom of online experience, the web. Libraries began to recognize the ubiquity of the web experience and the need to conform to it if they wanted attention, rather than, as in the online or electronic library, expecting users to make an effort to use library resources created by libraries themselves, which followed different conventions.

That shift is powerfully evident in the form of full-text searching. Look, for example, at Google Books. The indexing of that massive corpus of full-text book content is based upon the idea that users should be able to prospect the entire database by keywords. To have provided an equivalent search tool by manual indexing alone would have been impossible. An army of indexers typing catalogue cards for every word in every new book added to a vast collection would have produced a card index so huge it would have been completely unusable in any case. Google Books, some would say, is *itself* a digital library. The problem for institutional libraries is: should it be a part of *their* digital library? If so, how?

## Disintermediation

In the early days of the online versions of abstracting and indexing tools, such as the Science Citation Index and others, the searching was constrained by the high cost of personal computers and of computer bandwidth. This was why access to these online databases was offered to users only in mediated form. Researchers interested in a subject search for their PhD topic, or for a piece of research for which they were seeking or had recently received grant funding, would book a session with a librarian. The professional librarian running the search would ask the researcher to submit a search profile in advance, describing their topic and providing keywords for the librarian to use. This allowed the librarian to consider the online databases that would be most valuable for the search. When the researcher arrived, they would sit next to the librarian who would make the expensive network connection, log on to the various databases to be searched (to which the library would typically have a single subscription) and then run the search against each of them, with the researcher providing advice and answering questions about their topic for the librarian to be able to conduct the search. In a sense, the librarian had been turned into an instrument for the researcher to use, but it did not feel like that. It felt like an exercise in proving who the information expert was, and making sure that the answer to that question was, emphatically, the librarian.

Online searching was considered a skilled activity. Librarians attended training courses to learn how to search effectively. They would be taught how the Boolean operators worked, and what symbols to use to ensure that the AND, OR and NOT operations were effected in their search string. They learned how to save a search and re-run it later in a different database, or how to edit it to refine or expand the search. Thick manuals were produced by the database vendors and they included loose-leaf ring binders, with updatable pages as the databases developed, to incorporate new features or extend their coverage of the data.

The whole business was geared to the fact of scarcity. The available computing power was scarce. The network availability was scarce also, in the sense that it was expensive, as in those days the library would typically pay the database host system a per-minute charge to connect to their computers, as well as a charge, normally per record, for every item they 'downloaded' (or at least saved to a file to be printed later). The librarian was an optimizer in the process, trying to interpret the researcher's query and frame it appropriately for the databases being interrogated in the shortest possible space of time, incurring the smallest possible charge.

If we fast-forward to today's access environment, much has changed. Moore's Law has reduced the cost of computer power and ubiquitous network connectivity has liberated and democratized the wide-area network, or the aggregation of networks we now all share, to the point where mediation is unnecessary. Users are doing it for themselves. The online databases are still available, with many more having been added over the intervening quarter century. Abundance has replaced scarcity in a dramatic way. We have commercially licensed databases, and non-commercial open-access databases, many of which are created and run by university libraries themselves. We have databases of bibliographic records with links to other points on the web where the full articles or book chapters can be read for free or downloaded for a price that the library may already have paid, or one which the end-user can pay for themselves in an e-commerce transaction.

In a recent report from the University of Minnesota Library entitled *Discoverability*, Hanson concludes:

> "*Search, once one of the key skills and specialities of librarians, is now a daily activity for the vast majority of our users. Our users approach their research with an established history of search success that gives them confidence in their search skills.*" ([2], p 8)

This transformed environment has precipitated a crisis of confidence in some library professionals who saw themselves as highly skilled information diviners, and then discovered that their users didn't need them after all and they had been no more than a necessary evil. Librarians' control of the access to online scholarly literature was to be very much modified and reduced during the transition from scarcity to abundance. The impact has been that users have been empowered. Search is something we all do all the time. When we search we mostly don't think 'which directory?' or 'which index?'; we simply 'Google it'. This is convenient. We have cut out the need for selection, and made the search process much quicker and easier.

## From scarcity to abundance

In the online library, the explosion was very much a 'controlled explosion'. The library made online searching seem like a dark art; however, the art was not really in the query interpretation, but rather in the controlling of costs. Libraries are ultimately pragmatic solutions to the management of scarce resources. Back in the mid-1980s the scarcity was such that, despite the amazing power of computing to transform the data available and inherited from the print world, the access to that transformed, rich, exciting new corpus was highly rigid, selective and slow. There was a mismatch, and eventually the powers that made possible the explosion of information retrievability would become too strong for the scarcity model that had always operated in libraries.

Figure 1 shows what the impact in the explosion of retrievable information in universities has meant for libraries and the environments they manage. The changing environment has been simplified into four types: print; 'online' (by which is meant the early form of digital information managed by libraries, often called 'online' or 'electronic' information); digital (which is differentiated from 'online' mainly by implication: it implies that the digital form is end-to-end, e.g. metadata to full-text); and 'free web'. The free web environment is characterized by abundance. Anything can be put on it and accessed from it. That is not true of library environments, which are collections, acquired by selection, and so always relatively scarce. Library environments are also managed. The 'online library' has now moved completely into what we would call the 'digital library', but only a portion of the 'print library' has done the same. How that portion will change,

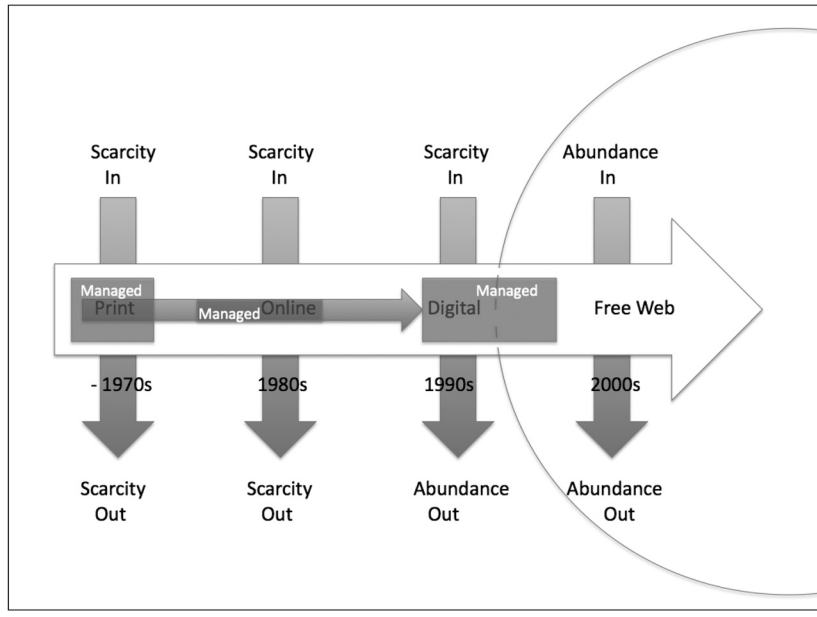


**Figure 1**

**Stages in academic library development from print to digital**

and over what timeframe, are estimates that have been notoriously difficult for librarians and others to predict.

One of the major changes to have impacted the library with the advent of the web world or the digital age is that they began to have to manage 'abundance out'. In the print library, and the online library, the environment was one of both 'scarcity in' and 'scarcity out'. By 'scarcity out' we mean that the managed resources are finite and possess a one-to-many relationship with their users. Libraries therefore require management practices, such as loan systems to circulate material fairly and equitably, that manage that scarcity. This remained true in the online library where access to the newly available computerized abstracting and indexing tools was mediated by the library in order to manage the high costs. When we move into the digital library, however, we see a major change. The library still manages 'scarcity in' as it still collects approved materials for its users, but those materials are increasingly acquired on the basis that they can be *used* abundantly, i.e. by as many library users as wish to use them simultaneously. This has not always been the case, indeed some digital resources are still making a transition to this model and libraries are required to license them by selecting a pre-set maximum number of simultaneous users. But more and more the practice is that they can be used campus-wide and the licence is purchased on that basis.

A growing proportion of what constitutes the digital library is provided from the abundant environment of the free web. The management of this has proved challenging for libraries. Their instinct initially was to assimilate these resources, to 'acquire' them for the library in the way they were used to doing with the materials they manage in these other environments. Even digital resources acquired under licence can be given library branding, although librarians often worry that this is not obvious enough. What libraries are now beginning to realise, however, is that the interface between the 'digital library', that they clearly manage, and that portion of the abundant free web that represents useful resources for their users, requires a different form of management from the one they have traditionally employed, in which material is assimilated (books are barcoded and class marked; journals are stamped; materials on the website are branded). The free web includes resources like Google Books and the Hathi Trust repository. It includes well-known themed repositories of research materials, such as arXiv, RePEc and PubMed Central. But it also includes large numbers of useful scholarly resources used, as well as created and, often, maintained, by small groups of researchers in sub-domains. This growing universe of freely available research materials cannot be assimilated and branded by libraries. That task would be too large and in any case futile.

## The loss of authority

What the user wants is what Google apparently provides: a source of indexed data that sits on the web with huge gravitational force, pulling data into the ambit of its indexers, crawlers and robots, and becoming a place to find the answer to everything. However, there are limits to Google. It does not index everything, and what it does not index it does not tell us. We cannot expose Google's parameters.

But it will give some satisfaction to almost every search request made of it. There is a lot of evidence from teachers in universities that students today start and stop their information searching with Google. *If it's not in Google, it doesn't exist.* They may sense that there are likely to be other sources that they could check, but Google throws back at them enough seemingly useful results that they can easily fill the time available to them to write their assignment by prospecting just those results and calling up the resources identified to use in their responses.

Google gratifies. The reference infrastructure that it has replaced did not yield gratification in the same way. However, it did provide authoritativeness. Every source searched, each index or abstracting service, enumerated its contents. It told users what journals it covered, or what were the parameters employed across the monographic literature. The infrastructure provided its own advanced *evidence* that a search was comprehensive. Google does not provide that evidence, and that is why serious researchers, those at levels above the undergraduate (and, arguably, within the undergraduate level, inasmuch as undergraduates are trainee researchers) require something more than Google.

But there is no simple 'either/or' option here, and this is a challenge to libraries both in respect of their roles as information literacy teachers, if they still maintain that role, and in the way they present their digital resources. A few years ago, librarians, and many academic teachers, would tell students that they should not use Google to search for information used in class assignments, or (later) that they should not use it once they had begun to write heavy-duty research dissertations, or that it was inappropriate for postgraduates, and so on. Yet these mass tools, Google, Wikipedia and others, have gone on getting better over time and through use. A student beginning to work on an assignment on August Strindberg, for example, would find that the Wikipedia entry on him is at least as good a place to start as most literary encylopedias. If that student then decides to try to find a copy of Strindberg's play *Miss Julie* on the web, they might begin with Google Books, but the search there is fruitless. They might look with more expectation in the literature database their university library buys on subscription from Chadwyck-Healy, Literature Online, but it has nothing in full-text either. If they were aware of the new, free web repository of digitized texts from several North American research libraries, the Hathi Trust, they would eventually be successful.

In this case, a free resource, not Google, would fulfil their search request in a way the service they might have expected would, could not. The Hathi Trust therefore takes its rightful place alongside Literature Online in the digital library that a researcher will recognize. Researchers may not think of it as a digital library. They may think of it as only one of a number of places on the web that are useful to them in supporting their research. This can trouble librarians with their urge to assimilate. If the Hathi Trust repository is good enough to meet researcher needs then the library should of course reference it from its website, so that those who are unaware may learn of its existence. To that extent, the library is still managing 'scarcity in' (as not all full-text repositories on the web would be deemed worth pointing to). However, it need not assimilate such services. Its mediation role is not, as in the days of the online library, co-extensive with the use of the service. It is now much less than that; it merely points from its website. And its users are

not clamouring for training in how to use these tools, as they are expert in the way websites work in general and can expect to find out as they go along.

Nevertheless the library's subject librarians should know about sources like Hathi, and Google Books, as well as Literature Online, in some detail. They should know about many more besides, down to a fine-grained sub-disciplinary level. Whether or not they teach students and researchers about them in formal settings, they must be knowledgeable enough to indicate their usefulness in individual and group interactions with users, and to understand and indicate their potential to serve as models for all disciplines and in contexts beyond the most obvious.

## Conclusion

In a world of abundant resources, libraries are still needed to manage scarcity, but the resources that are scarce have changed, and the meaning of scarcity has been expanded. When library books and journals were very scarce, they were managed as reference libraries. We still have such libraries (national libraries, for example) to manage scarcity in the sense of preservation copies of the nation's published output, and for rare or unique materials. The expansionist university libraries in the pre-web period of the last part of the 20th Century managed print materials that were becoming cheaper, and so turned most of their collections into circulating library collections, where there was a degree of tolerance of loss of materials as a trade-off for allowing users to manage their own access to the items in open-access stacks, and even, more recently, to manage their own borrowing and return functions.

With the web, however, we enter an age of apparent abundance. Yet there is still scarcity, for example in the materials that, although in digital form and conforming to the idiom of the web, are nonetheless commercially produced. These are sold to libraries on subscription, and the library still therefore manages these as a scarce resource, even though they behave in the same way as the abundant materials that researchers and students also want. The challenge in managing the digital library is to present an apparently abundant world of materials to users, which is in fact a hybrid environment of scarce and abundant resources, interlinked and often referencing each other across a 'pay wall'. Libraries have had to invent or commission clever bridging tools to allow this hybridity to operate invisibly, in order to allow users to experience search and fulfilment as though in an abundant world: abundant in scholarly resources and abundant in access provisions.

So while it may appear that the impact of the explosion of retrievable information upon universities has been the disappearance of the library, in fact what has disappeared has been the need *to go to* the library except (and importantly), as Daniel Greenstein noted, to make use of rare or unique materials or to find a place to study.

The management of 'scarcity out' gave libraries a reason to erect a wall between themselves and users. It allowed them to claim, often correctly, that on their side of the wall was not only the budget to pay for the materials that would be provided by the library, but also the expertise to make optimal use of them.

What was wrong, however, was their belief, or claim, that such expertise was unattainable except by the trained and initiated. The identification of access or provision with expertise has been lost due to the explosion of retrievable digital information and the shift to an expectation of 'abundance out'. Libraries still manage the budgets that provide the access, but they have lost their claim to scarce expertise in the finding and use of information. Users can do that for themselves, and they seem to prefer it that way.

## References

1. Steve Kolowich (2009) *Libraries of the Future,* Inside Higher Ed, 24 September 2009, http://www.insidehighered.com/news/2009/09/24/libraries
2. Hanson, C., Hessel, H., Barneson, J., Boudewyns, D., Fransen, J., Friedman-Shedlov, L., Hardy, M., Rose, C., Stelmasik, B. and Traill, S. (2009) *Discoverability Phase 1 Final Report*, 13 March 2009, http://purl.umn.edu/48258