

1

Digital futures in current contexts

In the paper-based environment, libraries and information centres have been the central links in the information chain. We are, however, in the midst of a profound transition resulting from the digitization of information, and all components of the information chain are in a state of flux. The parts played by authors, publishers, libraries and other information service providers are changing, and in many instances the boundaries which have demarcated the roles of these players have become blurred.

(Dorner, 2000, 15)

Just as the method of recording human progress shifted from the quill to the printing press 500 years ago, so is it now shifting from print to digital form. The library will continue to provide books and the printed record, but must now also deliver texts, images, and sounds to the personal computers of students, faculty and the public.

(Karin Wittenborg, www.lib.virginia.edu/dlbackstage/services.html)

Introduction

The worlds of both communication and the production of information are changing rapidly, and it is the convergence of these, and the consequent huge impact on libraries and library practice, that this book aims to address. In this introductory section, we examine the background changes in communication and information over the last 50 years, and the concomitant changes in libraries, as well as the changes in

the publishing industry. We give some basic technical definitions in order to elucidate some of the terms used elsewhere in the book.

This chapter will discuss the following issues:

- the information revolution in a wired world
- information explosion
- the nature of digital data
- storage and transmission of digital data
- developments in digital data creation
- printing and publishing
- changes in libraries
- digital libraries
- automating information retrieval
- the world wide web
- why the world wide web is not a digital library
- changing names for managing content
- unresolved issues.

Information revolution in a wired world

Over the last 50 years, the computer and communications revolution has changed radically the way many organizations do their business. According to Charles Jonscher (2000), we are now living in a wired world. With old-style twisted pair telephone wiring, co-axial cable, and optical fibre there are physical communication networks almost everywhere on the globe, and the places these do not reach can be covered by satellite. Business and military communication needs have promoted most of the telecommunications developments, and the rapid growth in mobile telephony, fax and e-mail have transformed business and financial transactions. But the greatest recent advance in communications technology was initially an academic development: the world wide web was invented by Tim Berners-Lee in the late 1980s in order to improve the storage and currency of electronic documents at CERN, the European particle physics laboratory. The rest, as they say, is history. But history has much to tell us about technological change and its effects on human society, which we would do well to note as a background to the changes affecting libraries that we are considering here. Humanity has always used

technology. The wheel is technology, a flint to light a fire is technology, anything that has been developed to do work so that human effort is easier is technology. In recent years, information and communication technologies have generally been referred to as ‘new’ or ‘high’ technologies – they are highly visible, and have not yet, despite their pervasiveness, become part of the natural infrastructure of society which, according to Borgman is invisible when it functions adequately and only becomes visible upon breakdown (2000, 19–20). ‘Technology’, as the computer scientist Bran Ferren memorably defined it, ‘is stuff that doesn’t work yet’ (Adams, 1999). Interestingly, Jonscher sees the development of information and communication technologies as a ‘curious turn’ that society took four decades ago away from mechanical and engineering developments and into micro-electronics, giving birth to the digital age (2000, 4).

The influence of new technologies on future developments is often wildly miscalculated, and human ingenuity is such that tools are often employed for purposes other than those for which they were designed. To give some examples, in 1876 Western Union suggested that the telephone had too many shortcomings to be seriously considered as a means of communication and was inherently of no value to them and in 1943, Thomas Watson, chairman of IBM, opined that there would probably be a world market for around five computers. More recently, Bill Gates, whose life work seems to be aimed at proving Watson wrong, famously stated that he couldn’t imagine that anyone would need more than 640K of memory in their computer. The internet was developed in the 1960s for a very restricted and specific purpose: the maintenance of communications in the USA in the event of a Soviet invasion: for Castells, it was ‘the electronic equivalent of the Maoist tactics of the dispersal of guerilla forces around a vast territory to counter an enemy’s might with versatility and knowledge of terrain’ (1996, 6). Who could have predicted the myriad uses it would be put to within less than three decades?

On the other hand, huge social changes have been predicted, mediated by technologies, which have not come about: supersonic aircraft, for instance, which in the 1970s, when Concorde was first designed, were supposed to revolutionize the speed of global travel, have remained a rare and expensive luxury. The infrastructure cost of supersonic flight has remained very high, while the cost of computing is reducing to that of television ownership owing to the social imperatives driving the technology

cost down. As Castells points out, 'technology does not determine society . . . technology is society and society cannot be understood or represented without its technological tools' (1996, 5). Society also has to be ready, both technically and psychologically, for major technological change to happen. For instance, all the underlying theories and algorithms for the stored-program computer were available by the 1850s, but mechanical limitations of the time meant that it could not be built until almost 100 years later (Swade, 2000). Progress is not a matter of smooth linear change, but 'a series of stable states, punctuated at rare intervals by major events that occur with great rapidity and help to establish the next stable era' (Gould, 1980, 226), a process characterized by Kuhn as 'paradigm shifts' (1970). Gleick (2000) suggests that these changes are happening with greater rapidity than ever before, which is probably why we currently feel change as a process of constant acceleration.

With the digital revolution, data and information can now be transmitted to all corners of the world, and that is significant for almost all humanity, and it is significant for libraries. But though many predict that we are reaching a halcyon period of cheap access for all, there are still political, cultural and financial issues that prevent this in many strata of society and many parts of the world. The digital divide exists and could further disadvantage the poor, the under-educated and those in developing countries as the better-off, the better educated and the economically developed race ahead into the digital future. Views on the democratizing nature of electronic networks vary wildly and we need to be cautious in our evaluation of these: for some we are on the verge of global utopia, an 'age of optimism' (Negroponte, 1995, 227), for others the internet 'continues to remain an expensive western toy' in a world where less than 2% of the population is connected to it (with London having more internet accounts than the whole of Africa) and where 80% of the population has never even made a telephone call (Taylor, 2001, 35). In southern Africa, 'for the ordinary citizens, who are in the majority, the financing of bread and butter needs takes a precedence over information' and the levels of charging 'impose a form of censorship' (Muswazi, 2000, 78). Indeed, we are more likely to see information equality being promoted by libraries than by other information organizations, as libraries, especially public libraries, have been the great information levellers for centuries, providing more (generally free) information to

their users than the latter could ever have by purchasing it. Libraries need to continue to provide this role in the digital age, even though many users are questioning the need for libraries and librarians in an era of massive free information resources available at the click of a mouse button. However, as we demonstrate in this volume, the information management skills of trained professionals are needed more than ever as we are overwhelmed by data of questionable provenance and unknown value. As Borgman (2000, 194) so cogently points out:

The claim that the Internet will replace libraries often is based on questionable assumptions. Three common misconceptions are that all useful information exists somewhere on the Internet, that information is available without cost, and that it can be found by anyone willing to spend enough time searching for it.

Implicit throughout this book is a refutation of all three of these misconceptions.

Information explosion

We feel as if there is more information around us and that information overload is a reality, but how much unique information is actually created each year and is it new or are we just being drowned in copies? Studies by Lyman and Varian at the University of California at Berkeley (available at www.sims.berkeley.edu/research/projects/how-much-info/) estimate that the world produces about 1.5 billion gigabytes of unique information per year. Apparently this equates to roughly 250 megabytes for every man, woman and child on Earth or the equivalent textual content of 250 books each. The study investigated the amount of *new* information stored in the four storage media (print, optical, film and magnetic), and asked how much storage would be required if it were all presented in digital form. Of course, this study does not tell us whether there is actually more information now than ten or 20 years ago although that seems extremely likely. We have to look to a tightly controlled intellectual community to see in more human terms the actuality of the information explosion as described in this insight into chemistry:

this year chemists will publish a hundred times as many papers than in 1901, when van't Hoff received the first chemistry Nobel prize.

(Schummer, 1999)

In 1961 the doubling period for the total amount of information in science was between 12 and 15 years (Price, 1961), by 1990 it was down to every six years (Braukmann and Pedras, 1990). Now some think it is as short as one to two years or less, although proving this for science is difficult now the internet is a major publishing point. The information explosion can also be measured in economic terms:

America's industrial output weighs about the same as it did 100 years ago, even though real GDP (Gross Domestic Product) is 20 times higher, reflecting the higher knowledge content of contemporary goods and services.

(Cronin, 1998, 3)

This explosion in information, services and resources, whether appropriate to the users' needs or not, consumes attention. Information has to be selected or discarded, read or not read, but it cannot readily be ignored. The actual downside of the information explosion is a deficit of attention, known more popularly as 'information overload'.

The nature of digital data

In order to understand the changes taking place in our information activities, it is vital to understand some of the underlying structures and principles of the digital world. All digital data, from whatever original it derives, has the same underlying structure, that of the 'bit' or the BInary digiT. A bit is an electronic impulse that can be represented by two states, 'on' or 'off', also written as '1' or '0'. A 'byte' consists of eight bits, and one byte represents one alphanumeric character. A ten-letter word, for example, would be ten bytes. Bits and bytes are linked together in chains of millions of electronic impulses; this is known as the 'bit stream'. A 'kilobyte' is 1024 bytes, and a 'megabyte' 1024 kilobytes. Digital images are represented by 'pixels' or picture elements - dots on the computer screen or printed on paper. Pixels can carry a range of values, but at the simplest level, one pixel equals one bit, and is represented in binary form

as ‘black’ (off) or ‘white’ (on). Images captured at this level are ‘bi-tonal’ – pure black and white. Images can also be represented as eight-bit images, which have 256 shades of either grey or colour, and 24-bit images, which have millions of colours – more than the eye can distinguish. The number of bits chosen to represent each pixel is known as the ‘bit depth’, and devices capable of displaying and printing images at high bit depths (36 or 48 bits) are now emerging.

Bit depth is the number of bits per pixel, ‘resolution’ is the number of pixels (or printed dots) per inch, known as ppi or dpi. The higher the resolution, the finer the texture of a digital image. The resolution of most computer screens is generally in the range of 75 to 150 pixels per inch. This is adequate for display purposes (unless the image needs to be enlarged on-screen to show fine detail), but visual content at this resolution is inadequate for printing (especially in colour), though images of black and white printed text or line art are often acceptable. High-density images of library originals (manuscripts, photographs, etc.) need to be captured in the range 300–600 ppi for print-quality output.

Almost any kind of information can be represented in these seemingly simple structures, as patterns of the most intricate complexity can be built up. Most primary sources held in libraries and other cultural institutions are capable of digital representation (anything that can be photographed can be digitized, though there are issues with the faithful representation of 3-D objects). When digital they are susceptible to manipulation, interrogation, transmission and cross-linking in ways that are beyond the capacity of analogue media. Creating an electronic photocopy of a plain page of text is not a sophisticated process compared with analogue technologies, but being able to then automatically recognize all the alphanumeric characters it contains, plus the structural layout and metadata elements, is sophisticated, and is only achievable in a digital environment. Alphanumeric symbols are the easiest objects to represent in digital form, and digital text has been around for as long as there have been stored-program computers. These symbols are also the most compact to store, an important factor at a time when capacity was limited and expensive. Early computers were good at processing symbols rapidly, but input, output and display of data were difficult. Input was via punched cards or tapes, output through printers or plotters with limited capabilities, and display was limited to ASCII symbols, displayed line by line,

often with only upper case available. Special fonts and characters were fearsomely difficult to represent, though some early pioneers made creative use of graph plotters for non-Roman character sets.

The storage and transmission of digital data

Early digital character representations were sparse in their storage needs, and early computer programs economical in their need for processing power. Storage and power have been increasing steadily over the past three decades, with the capacity of the processor chip doubling approximately every 24 months, much to the surprise of many experts (Jon-scher, 2000, 111-12), and although this trend cannot continue indefinitely for physical reasons, the immediate future will see further startling increases in computing speed. Increases in network capacity (also known as bandwidth) have also been huge, but rather less linear and predictable than increases in power because there are different factors that come into play here. Bandwidth is delivered through copper wiring (the installed telephone base still in most of the world) which has a relatively low capacity, through optical fibre cables, which have almost infinite capacity (Negroponte, 1995, 23), and through wireless transmission, which has variable capacity, given that portions of the airwaves have to be assigned with care to avoid interference of signals. Adding bandwidth can mean a process of digging up roads and laying cables to replace earlier wiring, a hugely expensive procedure, which must be funded by many different commercial and public bodies in most countries. Storage capabilities for digital data have also been growing exponentially: in the mid-1980s a 20Mb hard drive was regarded as massive storage, and 256Kb of RAM as sufficient memory. Now even laptop computers come supplied with 20Gb hard drives and 256Mb of RAM, which is 1000 times the capacity of machines of less than 20 years ago. But we need all this power and capacity to deal with the data that is coming our way in many media formats, as digital text, digital images, digital video and digital sound. An average page of digital text takes around 20Kb of storage space, which means that there can be 32,500 pages stored on an average 650Mb CD-ROM. High-quality professional digital cameras and scanners can potentially capture digital reproductions of the original in such detail as to require at least 300Mb of storage space, allowing just two

images to be stored on an average CD-ROM, and moving images and sound are even more hungry of space and power, though there are advantages in data compression that are helping to reduce the load. Data compression comes in two forms: 'lossless', which means that the compressed file loses no information in the compression and decompression processes, and 'lossy', where there is data loss in the processes. Lossy compression gives more compact results than lossless, resulting in smaller files to store or transport, but the derived images are of a lower quality – which may not matter if the files are for web delivery, but which is unacceptable for high-quality printing or long-term storage.

Increase of capacity and power in any of the component areas of digital processing and transmission are not just linear benefits: the relationship between demand and supply is complex and paradoxical. For example, a desktop computer would probably give excellent service for eight to ten years, but is likely to be replaced in most professional or even personal environments in two to four years (with four years being an average write-off period in many public and private institutions). The resale value is likely to be very low, if there is any value left in the machine at all. This is because new developments render hardware and software obsolete so rapidly. Compare this with other everyday machinery, the car for instance. There are people who change their cars every couple of years, for a variety of reasons to do with reliability, economy or prestige, but rarely because a particular model has become obsolete and will not run on certain roads, and there is certainly considerable resale value left in most cars. In the world of information supply, it is difficult to know whether the advances are driven by user demand or by the commercial imperatives of the suppliers of information and the suppliers of technology. Libraries are caught in the middle, having to purchase and provide ever more expensive and technologically demanding content to a wider range of users across a greater demographic and geographic area.

Developments in digital data creation

Digital text

Despite the difficulties of the early technology, some scholars recognized the value of the computational manipulation of textual materials imme-

diately. Text had to be painstakingly entered on punched cards or tape, but the benefits appeared so great in terms of retrieval and analysis that they persisted. Father Roberto Busa, for instance, formulated the idea of automated linguistic analysis of text in the years 1942–6, and started working with IBM in New York in 1947. He produced more than six million punched cards for his edition of the works of Thomas Aquinas, and in 1992 the first edition of his CD-ROM was published (Busa, 1998). The classical scholar Anthony Kenny wrote a book on statistics in the early 1980s so that other scholars could learn to manipulate electronic text for scholarly purposes (Kenny, 1982). In the 1980s a new technology appeared in the form of the Kurzweil Data Entry Machine (KDEM), which was developed to provide texts for the blind. It used optical character recognition (OCR) to create electronic text for printing on Braille printers or (eventually) conversion to sound using text-to-speech processing. The KDEM was extremely expensive (in the tens of thousands of dollars), but it could be trained to recognize different typefaces and fonts, and it was relatively accurate – probably as accurate as modern OCR packages. Other text conversion needs could, of course, be satisfied using the KDEM, and large corpora of digital text were created for many purposes. The Oxford Text Archive, for instance, gathered many of the texts that it still supplies as output from the KDEM service operated nationally by Oxford University until the early 1990s, and these have been used all over the world as the basis for digital text collections, at the Universities of Michigan and Virginia in the USA, for instance. Digital text is now everywhere, as output from word-processors, publishers and OCR engines. Interestingly, OCR is faster and cheaper than it was in the 1980s, but it is still not as accurate as it needs to be to replace rekeying and typesetting, especially if the materials are pre-20th century. Many library projects use OCR to capture bodies of printed text; others, where the text is too difficult or a higher level of accuracy is needed, use rekeying. However, there are some new developments in the processing of text that greatly improve retrieval from inaccurate OCR; these are discussed in Chapter 2 ‘Why digitize?’.

For maximum usefulness, digital text needs more than representations of alphanumeric symbols on the printed page; it also needs metadata to record other information about the textual object from which it derived. Markup languages such as the Standard Generalized Markup Language (SGML), described in more detail in Chapter 5 ‘Resource discovery,

description and use', define textual metadata and specify complex schemas of standard metadata tags to inform all the processes to which digital text might be subject: description, retrieval, preservation, print output, and so on.

Hypertext, multimedia and digital images

While early document-processing technologies generally operated on strings of linear text, the possibilities of hypertextual linking within documents were recognized too, though this was initially difficult to implement on a computer. Despite the claims of many modern gurus of cybernetic hypertext theory, written text is not linear but can be interlinked, interwoven and annotated on the printed page in highly complex structures not susceptible of computational representation by the first generations of hardware and software. From the mid-1980s, better screen technology, more processing power, the development of the mouse and the advent of the Graphical User Interface (GUI) have transformed the ability of the computer to represent and link documentary objects. In 1996, for example, the first CD-ROM of the Canterbury Tales project appeared: the Wife of Bath's Prologue, published by Cambridge University Press. This presents transcriptions of 58 manuscripts and early printed versions of the work, and generates almost two million hypertext links (Robinson, 1996).

In the last five years, digital camera technology has developed to the extent that digital images can be captured that equal or even exceed large-format analogue photographic reproductions. These have disadvantages, however: the cameras are expensive, and the file sizes huge, which causes problems for storage and delivery. However, from valuable originals, large archive images can be captured and stored for the long term, with lower-quality derivatives being delivered for viewing and printing. This is acceptable for most uses, and means that good-quality images can be integrated with other media for a more complete user experience. As of 2001, the British Library has produced a high-quality digital facsimile of the 15th-century Sherborne Missal that is on display in its galleries on the largest touch screen in the UK. The unique feature of this resource is that the pages can be turned by hand, and it is possible to zoom in at any point on the page at the touch of a finger on the screen. High-quality sound

reproduction accompanies the images, allowing users to hear the religious offices that make up the text being sung by a monastic choir. Now documents can be represented by high-quality images, with underlying searchable text, and with annotations in text, sound or video. We give some more examples of the digitization of complex documentary collections in Chapter 2 ‘Why digitize?’. Cameras are being developed that can capture 3-D images, and there are sophisticated capture devices for digital sound and video, deriving in part from the huge growth in these formats in the entertainment industry.

It is, of course, possible to treat the printed page as an image. Capturing digital images from printed pages has the advantage that the page will appear on the screen exactly as it was in the original. The disadvantage of this presentation is that the search and retrieval functions are lost, but a combination of both image and text representation means that documents can be represented by images, with underlying searchable text and with annotations. We give some more examples in Chapter 2 ‘Why digitize?’.

Printing and publishing

There is a rhetoric that suggests that we are moving rapidly from print to digital media: this is not borne out by evidence from the publishing industry and from copyright libraries (Leonhardt, 2000, 123). The death of the book has been predicted with every new communication or entertainment technology, the telegraph, recorded speech, film, television and the internet, but the book seems to be thriving. At the beginning of the 21st century, the UK copyright libraries, for instance, are receiving more print material than ever before (in 1999 the figure was around 105,000 items for the British Library alone), and this is without the materials that they purchase. Production of digital data is certainly on the increase, but it does not seem to be accompanied by a concomitant diminution in the printed output, though it is changing both the publishing industry and the world of libraries. In the 15th century, the introduction of the printing press industrialized the production of books, but did not do away with handwriting. What it did do was change book production from a cottage industry carried out in monasteries to a commercial industry, and incidentally resulted in a huge

loss of authority on the part of the Christian Church in Europe. The digital revolution is often likened to the print revolution, but it is still rather too early to tell whether the effects will be as far-reaching and transformative.

In the 14th century, monastic production could not keep pace with the demand for the written word, despite large-scale copying throughout Europe. There was a huge potential market for multiple copies of books, and there was also a demand for greater accuracy than copying was supplying. Chaucer's plaintive appeal to his scribe, Adam, exemplifies the problems of careless and inaccurate copying that plagued authors and readers alike:

So ofte adaye I mot thy werke renewe,
It to correcte and eke to rubbe and scrape,
And al is thorough thy negligence and rape.

(Chaucer, 1988, 650)

Every day I have to redo your work, correcting it and scraping away the errors, all because of your negligence and carelessness.

Johannes Gutenberg, born sometime in the last decade of the 14th century, and trained as a goldsmith, produced the first book in Europe printed from movable cast type, the so-called 42-line Bible. Despite popular misconceptions, Gutenberg did not invent this method of printing, which was known in China and Korea from around the 11th century. What he did was to combine

the technology of the goldsmith's punch with that of the winepress. The result was the printing press – a machine that combined flexibility, rapidity, and economy to allow the production of books that the increasingly literate, increasingly numerous European city dwellers could afford to buy and read.

(Lerner, 1998, 96)

The 42-line Bible was produced in around 1455.

Hand-set movable type not only produced books faster than handwriting, it produced them in multiple copies. While with early printed books it is not quite true to say that every copy is the same, the differ-

ences are small compared with the differences between manuscript copies. And printed books can be produced in the hundreds and thousands, even millions, which is the significant advance. Evidence for the instant huge potential market for multiple copies can be found, according to Kilgour, in the estimates of the numbers of books printed in the last third of the 15th century, that is in the 30 years after the death of Gutenberg in 1568. These estimates suggest that there were some 12 to 20 million books printed – more than the number of all manuscripts produced in medieval Europe up to that time (Kilgour, 1998, 82).

More books call for more readers, and so literacy rates rose rapidly in the next three centuries. Paper was made from linen rag during this period, but in the Napoleonic Wars rag was needed to make bandages, so wood-pulp paper processes were developed. Wood was plentiful and the process cheap, which meant that print output could increase. This happened in several ways: more publications were printed, and works were often longer, hence the growth in the multi-volume work of fiction in the 19th century. Periodical publications were produced which appeared monthly, then weekly, then daily and the mass media were born.

Early printing presses were wooden and were hand-operated, and the type was hand-set. This remained the case until the beginning of the 19th century, when things changed rapidly with one of those ‘paradigm shifts’ discussed at the beginning of this chapter. Printing-press technology moved to iron presses, operated by steam, to mechanical type, to hot-metal type to phototypesetting and then to computer typesetting by the 1970s. Publishing began to become a separate industry from printing in the 18th century, and grew rapidly throughout the 19th century with the increasing mechanization of the processes, and the further growth of literacy rates and the demands of a reading public. ‘The total book production of the nineteenth century exceeded that of the eighteenth by 440%’ (Kilgour, 1998, 112), and many of the publishing giants that still control the industry (though in very different configurations) came into being during this period.

Computer typesetting brought an unanticipated by-product: electronic text. This was not always kept initially, and when it was, the typesetting codes used for marking up the text rendered it largely unusable for any other process. However, standardization of markup languages (discussed in more detail in Chapter 5 ‘Resource discovery, description and use’) has

meant that it is now possible to use electronic text for purposes other than printing, and many publishers have embraced electronic publishing in the last five years, especially in the production of journals. We discuss the implications for libraries of the acquisition of large volumes of published electronic books and journals in Chapter 3 ‘Developing collections in the digital world’, and we also look at some of the models of electronic publication that are changing the face of both the print and the digital worlds.

Changes in libraries

With the revolution in printing and the birth of the publishing industry came new developments in libraries. In the Middle Ages in Europe, libraries were either monastic institutions holding relatively few, very valuable books (and many of them were chained libraries, that is the books were chained to the shelves to prevent theft) or belonged to private individuals. Many of the great libraries of today are based upon some of the collections of these monasteries and individuals. The collections of Archbishop Matthew Parker are vital to the library (named after him) at Corpus Christi College Cambridge; the British Library’s founding collections were those of Sir Hans Sloane, Sir Robert Cotton, Edward and Robert Harley, earls of Oxford, and the Royal Collection given by George II in 1757. When the British army invaded the city of Washington in 1814 and burned the Capitol, including the 3000-volume Library of Congress, Thomas Jefferson sold his personal library to the Congress to ‘recommence’ its library. The purchase of Jefferson’s 6487 volumes for \$23,940 was approved in 1815. In the 18th century, circulation libraries and subscription libraries started to appear to satisfy the demands of the larger reading public for cheap access to print, and the 19th century saw the rise of the great public libraries.

Mechanization

Librarians have always sought to mechanize routine processes as much as possible, and so were early adopters of computer technologies. Database programs were developed early in the history of computing for use in stock control, payroll systems and other commercial activities.

The adoption of administrative systems for catalogue record creation

was an early example of the use of computerized processes in libraries. This started as merely a means of producing, for manual filing, printed catalogue cards or slips, which would be referred to by library staff and users as a means of finding library resources. To extend the duplication of catalogues and thus the number of user access points to the catalogue, microformats were used, 'enabling multiple copies of the catalogue to be distributed for the first time and new sequences, such as title catalogues, to be introduced' (Brophy, 2001, 106). Alongside this new capability to share catalogues, there started co-operative efforts to share the cataloguing load and to benefit from aggregating cataloguing effort across multiple libraries. Because bibliographic description is such a highly structured construct, the computerization of the catalogue was a significant and inevitable next step from these early mechanization initiatives. This was not the computerized catalogue as we know it now in the guise of the OPAC (Online Public Access Catalogue), but a straightforward listing of the library's resources with no links to borrower records or to external resources.

Integrating computerized functions

In tandem with the development of the computerized catalogue came the move to automate circulation functions. Initially, this was no more than associating a borrower number with a book accession number while the full records of each were kept on separate computer database or paper systems. Even these most basic functions were only available to the largest libraries because of the great expense involved. However, this was to change as computing technology became more widespread and software developed the capability to create connections between different parts of databases and to integrate functions in a primitive fashion. As soon as it was possible to get the circulation system to 'talk' to the catalogue, it was possible to automate some library functions, such as overdue notices. As important as the mechanization of routine staff-intensive functions, was the means it gave library managers to measure and assess borrower activity in great detail for the first time through reports generated from the computer logs and databases. This knowledge greatly enhanced acquisitions and stock management strategies and has helped librarians to cope proactively with changes in funding structures, bor-

rower requirements and the user community, and with the accelerating expansion in information resources.

Changes in computing and networking

At the start of the 1980s, automated library functions were generally being achieved through mainframe computers with dumb terminals in large institutions, and it was not until the spread of personal computing that computerization could be seriously afforded by small libraries such as those in schools, smaller businesses, hospitals or other small specialized organizations. The concentration of computer activity in one central processing point also limited those libraries with geographically distant outstations such as public libraries or large universities, and often the computing resources could only be made available in the main library location. It was only with the inception of personal computers, local- and wide-area networks, and the movement from centralized to distributed client-server processing that libraries of all types were able to utilize fully the power of computing technology to automate functions, share information and resources, generate management reports and electronically link libraries together. From the perspective of the library user, they were experiencing computers in the library for the first time not as something which librarians used to facilitate library management, but as a direct point of service to use independently of the library staff.

By the early 1990s, almost all library functions were supported by automation in some way. Functions such as cataloguing, acquisitions, journals, circulation, interlibrary loans, financial control, stock management and user details were all integrated and automated to some extent. This period also saw a large growth in the electronic communication of data from one library to another. Librarians had for many years accessed remote database services for access to reference information and bibliographic data, using a dial-up connection to a local computer which would route information to the host machine, possibly based somewhere else in Europe or America. This was an early form of internetwork access, charging heavy fees not just on connect time, but on a per-item-retrieved basis. In most libraries the end-user could not be given direct access because of the high cost of use (with legal and pharmaceutical-based libraries sometimes the exception) and the librarian was required as a skilled

intermediary, and also as a buffer against over-expenditure. As wide-area networks in the USA, UK and elsewhere developed, the emphasis on connect time was relaxed as annual subscriptions with unlimited access were introduced by suppliers. CD-ROM versions of databases for personal computers also relaxed the need to limit end-user access. Wider use of networking enabled libraries to share more data in the form of interlibrary loans, electronic document delivery, electronic mail and also electronic ordering functions, also known as EDI or Electronic Data Interchange.

Electronic Data Interchange

Electronic Data Interchange (EDI) refers to the computer exchange of business information using a standardized data format. Standardized EDI messages are based on common business documents such as purchase orders, invoices and delivery notes, which are interchanged between computer systems without human intervention or interpretation. It has been in use since the 1970s, but it was not until the late 1980s and early 1990s that any attempt was made to move away from proprietary systems to open standards.

Adopting EDI can eliminate the mailing of paper documentation and the manual processing of quotations, purchase orders, invoices, shipping documents, customs documents, and other business transactions. Because the data is processed and stored automatically, tasks such as re-keying data and printing purchase orders and invoices are eliminated.

(Tallim and Zeeman, 1995)

The use of EDI has been as significant a development for libraries as for the business world. The integration of the cataloguing and ordering process, plus the ability to chase orders that have not arrived on time automatically, speeds up acquisitions and gets resources to the user more quickly on their arrival. Automated financial controls can be pre-set to generate purchase orders and fiscal reports, and also to limit certain types of expenditure to defined budgetary plans. Librarians, as early adopters of EDI, have been among the first to partake in e-commerce and thus understand e-commerce's trials and tribulations but also its great advantages to the management of library services.

Co-operation to mutual advantage

As far back as 1902 we can trace how basic technology was used and the results shared by libraries in a co-operative way. In that year the Library of Congress began selling printed cards that in effect ‘made that library the centralized cataloguing agency for thousands of American libraries’ (Lerner, 1998, 194). In computing terms we have to wait until 1971 for the same level of resource sharing and economy, when OCLC introduced an online shared cataloguing system for libraries (available at www.oclc.org/). Today this co-operative catalogue is available to libraries in 76 countries and territories around the world, with over 46 million cataloguing records available, as well as associated interlibrary loan and reference database services. In the UK, the Birmingham Libraries Co-operative Mechanization Project (BLCMP) was established in 1969 as a joint venture between libraries in Birmingham. BLCMP was created to realize the benefits of a shared approach to library computing and built a large shared database of over 19 million bibliographic records (available at www.blcmp.org.uk/). The Research Libraries Group (RLG) also provides extensive shared resources through a set of online catalogues that offer millions of records describing materials created around the world (available at www.rlg.org/libres.html).

In many areas, libraries are showing intensive co-operation and sharing of the fruits of computing technology. NORDINFO (the Nordic Council for Scientific Information), which celebrated its 25th anniversary in 2001, was founded as a bridge between Nordic research libraries and the growing information and documentation sector. The general objective set for the coming years is to work towards what has been called ‘The Nordic Electronic Research Library’.

Another significant example of sharing for mutual benefit is the important realm of interlibrary loans. No single library holds all published information and most cannot hold even a sizeable sliver of it, but the librarian’s role is clearly to provide access to the complete body. The only sensible solution is to share access to the resources through interlibrary loan. All the above co-operatives support this and local co-operatives for interlending exist all over the world at regional and national levels. Most of this is now being managed or documents delivered via electronic means. The British Library Document Supply Centre, for instance, received 34% of requests by postal means in 1993–4 but

now receives less than 13% by post and the remainder by electronic means (Maynard, 2000).

Digital libraries

The digital library, the electronic library (generally taken to be synonymous with the digital library), the virtual library, the hybrid library, the library without walls are all concepts that librarians seem to be dealing with all the time. What do they mean? Do they mean the same to everyone who uses the terms? Do they all mean the same thing? Do we all mean the same thing when we talk about a library? From its original etymological meaning of a 'collection of books', a library can mean a collection of almost anything in modern parlance: software routines, for instance. Every library is different, every digital library is different, and different players are advancing many definitions for the digital library. Arms (2000, 2), for instance, defines a digital library as:

a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network. A crucial part of this definition is that the information is managed.

For the Digital Library Federation in the USA:

Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

(Greenstein, 2000)

Borgman (2000) devotes a whole chapter to trying to define a digital library (called 'Is it digital or is it a library?') and concludes that (as is to be expected) the term has multiple meanings, and, for her, these cluster around two themes:

From a research perspective, digital libraries are content collected and organized on behalf of user communities. From a library-practice per-

spective, digital libraries are institutions or organizations that provide information services in digital forms. (51)

Some prefer to work with the concept of the 'hybrid' library, as they believe that this more closely describes the reality of libraries, which have always been hybrid. The digital is one more (albeit very different) format that librarians have to deal with in a multi-format environment, where for many years non-documentary mixed-media objects, both analogue and digital, have been growing in importance, and where technology for access has had to be provided. Most libraries can supply microfilm and microfiche readers, video players, audio tape and CD players, as well as terminals for access to digital resources. However, there are often strategic reasons for defining activities as coming under a digital libraries heading – the possibility of securing funding, for instance, from programmes which are defined as digital library initiatives.

Hybrid libraries are, according to Rusbridge (1998):

Designed to bring a range of technologies from different sources together in the context of a working library, and also to begin to explore integrated systems and services in both the electronic and print environments.

They exist 'on the continuum between the conventional and digital library, where electronic and paper-based information sources are used alongside each other' (Pinfield et al., 1998, 3). The concepts of virtual libraries or libraries without walls reflect the integrative possibilities inherent in the digital: if significant library collections are digital, then the confines of space no longer define boundaries upon information. Virtual collections from many different sources can be assembled and accessed from anywhere, without the user even knowing where the sources reside, and personal virtual collections can be built to serve many purposes. Hence a virtual library could potentially be enormous, linking huge collections from all around the world together, or it could be very small, being the personal digital collection of one individual.

Digital libraries are, like any so-called revolutionary change, a development of a whole range of underlying theories and technologies that have come together to create a paradigm shift. The speed of recent developments has taken some librarians by surprise, especially the exponen-

tial growth of the amount of digital data available, but they are understandable if we look at some of the precursors that have led to the current trends: new developments rarely spring fully formed from the ether.

For the purpose of the present work, our understanding of digital libraries is somewhat eclectic. There are many different kinds of digital libraries creating, delivering and preserving digital objects that derive from many different formats of underlying data, and it is very difficult to formulate a definition that encapsulates all these. We would, however, like to propose some principles that we think perhaps characterize something as a digital library rather than any other kind of digital collection. These derive in part from the definitions above and in part from our own experience. They are:

- 1 A digital library is a managed collection of digital objects.
- 2 The digital objects are created or collected according to principles of collection development.
- 3 The digital objects are made available in a cohesive manner, supported by services necessary to allow users to retrieve and exploit the resources just as they would any other library materials.
- 4 The digital objects are treated as long-term stable resources and appropriate processes are applied to them to ensure their quality and survivability.

Automating information retrieval

Vannevar Bush, one of Roosevelt's advisers in World War II is generally credited with being the first thinker to suggest mechanical and electronic means of dealing with complex information, and of finding paths through information universes that could be marked for others to follow. Bush was grappling with the problem that has beset humanity since the invention of printing, and possibly before. How, when so much information is available, does one remember what has been read, make connections between facts and ideas, and store and retrieve personal data? Pre-literate humanity was capable of prodigious feats of memory, but the total amount of knowledge it was necessary to absorb was orders of magnitude smaller than it is now. When writing became widespread, personal notebooks and handbooks were used as *aides-mémoires*, but these

could handle only a fraction of the data that even one individual had to cope with by the 1940s. In his seminal article, 'As we may think', published in the *Atlantic Monthly* magazine (1945), Bush conceptualized 'thinking machines' that would help the modern world make sense of the information explosion, and he proposed that an automated 'memory extender' be employed, a device he called a 'MemEx' in which an individual could store all his or her books, records and communications. This would be mechanized to enable rapid and flexible consultation. Interestingly, given developments in digital technology at that period, the MemEx was essentially an analogue machine and used microfilm for information storage with a mechanical linking process. The machine was never built by Bush, but drawings of it exist, and a simulation of it was produced some years ago by Nyce and Kahn (1991).

For some reason, Bush's article caught the popular imagination: *Life* magazine published an illustrated version entitled 'A top US scientist foresees a possible future in which man-made machines will start to think', and it was summarized in *Time* with the title 'A machine that thinks', and Bush has even been hailed as a 'Father of Information Science' (Buckland, 1992, 284). He is certainly thought of as the father of hypertext, even though the word itself was not coined until the 1960s. He is, however, rarely examined in relation to the developments he was drawing on, according to Buckland. 'Little attention, in contrast, has been paid to Bush's MemEx in relation to its *own* context: the visions and technological developments of information retrieval in the 1930s' (1992, 284). Buckland (1992) and Rayward (1994) attribute many of the developments underpinning Bush's theoretical achievements to Paul Otlet, the Belgian bibliographer who had helped develop the Universal Decimal Classification, and to Emanuel Goldberg. Indeed, for Buckland 'Bush's understanding of information retrieval was severely incomplete' (1992, 285).

Emanuel Goldberg developed very high-resolution microfilm, and also the technology underlying microdots, in the 1920s. In 1932, he wrote a paper describing the design of a microfilm selector using a photoelectric cell, which Buckland sees as 'the first paper on electronic document retrieval' (1992, 288). In 1934, Paul Otlet published his *Traité de Documentation* which 'is perhaps the first systematic, modern discussion of general problems of organizing information . . . one of the first infor-

mation science textbooks' (Rayward, 1994, 237-8). Otlet saw huge possibilities in new forms of information recording and transmission: he was aware of the information and communication possibilities of telegraph, telephone, radio, television, cinema and sound recordings. For Otlet, 'the book is only a means to an end. Other means exist and as gradually they become more effective than the book, they are substituted for it' (Otlet, 1934, quoted in Rayward, 1994, 244). Otlet visualized new kinds of work desks where there would be no documents, only a screen and a telephone with access to documents on film which could be called up at will. Otlet also formulated some new organizational principles which would hold all the documents in a linked relationship: a Universal Network for Information and Documentation which would connect centres of production, distribution and use regardless of subject matter or place (Rayward, 1994, 246) – a development that uncannily presages the world wide web.

Though Otlet and Goldberg may have formulated the underlying theories upon which hypertext and information retrieval, and therefore digital library developments, came to be based, it was Bush's legacy that inspired later thinkers and developers. In 1962 Douglas Engelbart (an early follower of Bush) started work on the Augment project, which aimed to produce tools to aid human capabilities and productivity. He also was concerned that the information explosion meant that workers were having problems dealing with all the knowledge they needed to perform even relatively simple tasks, and Augment aimed to increase human capacity by the sharing of knowledge and information. His NLS (oN-Line System) allowed researchers on the project access to all stored working papers in a shared 'journal', which eventually had over 100,000 items in it, and was one of the largest early digital library systems. Engelbart is also credited with the invention of pointing devices, in particular the mouse in 1968.

The mid-1960s saw other pioneers conceiving of grand schemes that at the time seemed so innovative as to be impossible, but now are coming to realization. Ted Nelson designed his Xanadu system in 1965, in which all the books in all the world would be 'deeply intertwined' (in his words – Nelson incidentally coined the word hypertext). Nelson also tackled the problems of copyrights and payments by proposing that there should be electronic copyright management systems that would

keep track of what everyone everywhere was accessing, and charge accordingly through micro-payments (McKnight, Dillon and Richardson, 1991). Impossible to implement at the time (Xanadu was described by Wolf (1995) as ‘the longest-running vapourware story in the history of the computer industry’), these ideas are now commonplace, but it took Tim Berners-Lee to put them all together in the late 1980s.

The world wide web

The problem that the world wide web was initially created to solve was document management. Berners-Lee was employed by CERN in the 1980s to devise a system that would allow the organization to keep track of all versions of all the documents that its employees were creating, editing and exchanging constantly. This is the same problem that Engelbart was dealing with 20 years earlier with NLS and Bush 50 years earlier with the MemEx. The web is based upon a relatively simple set of concepts: it relies on there being an underlying network of networks that can connect any computer to any other computer (which do not have to be of the same type) provided connections can be routed to allow them to exchange information. Given this infrastructure, what Berners-Lee did was implement the information management, storage and exchange goals of Goldberg, Otlet, Bush and Engelbart before him, drawing, consciously or unconsciously, upon all their underlying ideas. The key to the development of the web was the complex linking potential of hypertext, and Berners-Lee added to this Hypertext Markup Language (HTML), which allowed all text for the web to be encoded using the same system, the Hypertext Transfer Protocol (HTTP), to specify how information exchange between machines should be handled, and the Uniform Resource Location (URL), which specified addresses for documents. Despite huge developments in web technologies over the last ten years, the underlying principles are relatively unchanged. These developments are discussed further in Chapter 5 ‘Resource discovery, description and use’.

Why the world wide web is not a digital library

The world wide web has many of the features of a digital library, and if the web did not exist our conception of digital libraries would be very

different. The web is undoubtedly the means via which most digital libraries are accessed, but it is not a digital library itself as it lacks those characteristics we suggested would define digital libraries. It is not a managed environment, it has no collection development principles and most significant of all, the digital objects are not perceived as having durable value – though many of them do. Indeed, one of the issues being tackled in the digital library world is the vexing one of how to guarantee a record of the information available on the web for future generations. Much of the web is ephemeral information: advertising, personal web pages, announcements, etc., which come and go with great speed. In the past, this sort of information was available on paper, in handbills, newspapers and many other forms of prints. Much of it survives, albeit haphazardly. Accidental survival of web resources is unlikely. We deal with these issues in more detail in Chapters 5 and 8 ‘Resource discovery, description and use’ and ‘Preservation’.

Changing names for managing content

The integration of technology into the natural working practices of many organizations has led to a number of management initiatives that continue to have a distinct impact upon librarians. These are now drawn together under the umbrella term of content management, but have gone by many names before:

The names may change, but in many ways the story is the same. Over the past several years, we have seen a quick and accelerating shift from one term to another that attempts to name the technology responsible for creating, updating, managing, and distributing material in many forms . . . Call it a ‘document’, ‘knowledge’, or ‘content’, the problem set that was identified years ago is, at its core, the same. There is simply more of everything – more core material, more forms of it, and more ways to distribute it.

(Trippe, 2001, 22)

It is in these corporate environments and management methods that many of the digitization, storage and access solutions were first implemented and developed and then put to wider use. Librarians may be said to have managed containers of knowledge content in the past, but the

growth in computing use from the 1960s onwards propelled librarians and other information workers into the management of content. Megill (1997) points out that the information an organization needs to keep for re-use, that which is worth sharing, managing and preserving to function effectively, is the 'corporate memory'. It is the job of 'information managers' to keep, store and release this information in a timely fashion and as Lerner points out, special libraries 'exist solely to make expensive professional workers more effective at what they do' (1998, 182). Technology was introduced to try to resolve some of the difficulties inherent in corporate environments where 80% of corporate information was to be found in documents and e-mail, rather than structured database records (Megill, 1997, 7) and thus was difficult to retrieve. This was termed 'information management' or 'document management' as it drew together people from wide professional backgrounds, including librarians, record managers, archivists and computer scientists. It was the cutting edge for a time, but soon it became noticeable that the ability to find a document, drawing or data was not considered cutting edge enough to fulfil the corporate desire for flexibility, productivity and control. By putting technological solutions foremost it had been forgotten that information is internalized and interpreted by people: the 'knowledge' element, that which we know, was missing. It is interesting to note that the immediate response to this was to implement even more technology under the heading 'knowledge management'.

Knowledge management (KM) is often discarded as nothing more than a buzz term for something that was already in place. Brophy describes it accurately as 'better thought of as the process of engineering conditions under which knowledge transfer and utilization happen' (2001, 37). One would think this would signal a shift from technology towards people (such as the focus of community provided by libraries), but KM is renowned for implementing intranet solutions with complex knowledge databases to allow workers to share their information for the benefit of the whole organization. Of course, to transfer knowledge is more difficult than to record it, and learning (the process of acquiring knowledge) is not inherent to many of the technology-based solutions on offer.

The importance of people as creators and carriers of knowledge is forcing organizations to realize that knowledge lies less in its databases than in its people. It's been said, for example, that if NASA wanted to go to the moon again, it would have to start from scratch, having lost not the data, but the human expertise that took them there last time.

(Brown and Duguid, 2000, 122)

KM is fast becoming outdated as a concept and so in this new century we are returning to information in the term 'content management'. This is such a wide term, encompassing a broad and complex field, because as Gilbane points out 'everyone needs to manage content, but the similarity ends there' (Trippe, 2001, 22). Undoubtedly, modern libraries are in the business of content management and so how librarians and libraries will fit into this new management 'reality' is discussed in Chapters 3 and 9 'Developing collections in the digital world' and 'Digital librarians: new roles for the Information Age'.

Conclusion

Librarians, as we have shown here, are always at the forefront of the latest technologies in order to find new ways to optimize the management of libraries and resources, and to provide improved services. There are some outstanding problems with which libraries are still struggling that have to some degree been exacerbated rather than resolved by technology. Interoperability has been a significant technical and political problem for libraries ever since computing was first used to share data. (Interoperability is discussed in detail in Chapter 6 'Developing and designing systems for sharing digital resources'.) It is probably fair to say that all attempts at full interoperability between libraries, systems and standards, and between communities have not yet succeeded, and are unlikely to succeed. Even assuming technical hurdles can be overcome, there are also the political issues of control, resourcing, legal frameworks, regional, national and international community differences, and the traditional boundaries of different cultural sectors such as libraries, schools, museums, galleries and archives. The goal of the world-wide global library is probably unreachable while the issue of interoperability still remains as the biggest sticking point of all.

The other unresolved strategic issues revolve around money, infrastructure, scalability and sustainability. Computing in libraries is no longer showing the immediate cost savings in return for investment that were delivered in the 1980s and early 1990s. So far in this chapter we have discussed the way that computerizing routine tasks has saved time and money, while enhancing services and freeing library staff to do more management and less paper chasing. The situation now is that these developments are taken for granted in many organizations and future developments will not necessarily instantly save staff time or reduce costs. The real benefits from technology now for libraries are improving resources and services, not replacing the human factor. This is discussed in more detail in Chapter 4 ‘The economic factors’, where we examine the financing of technology, and in Chapter 9 ‘Digital librarians: new roles for the Information Age’, where we evaluate the changing library profession. Of course, the issues of sustainability and scale become paramount once significant investment has been made and there is an undercurrent of dissatisfaction about the sustainability and scalability of digital technology. Issues of preservation and continuing access to resources are discussed in Chapter 8 ‘Preservation’ in particular.