Digital libraries and archiving knowledge: some critical questions

Peter Johan Lor[1]
IFLA, P O Box 95312, 2509CH The Hague, The Netherlands
peter.lor@ifla.org

and

Department of Information Science, University of Pretoria, Pretoria 0002, South Africa.

Received: 10th September 2007
Accepted: 7th June 2008

Over millennia librarians have striven for universality: complete control of all recorded knowledge, if not through ownership then through bibliographic organisation and systems for universal availability and access. Modern digital technologies offer new possibilities of achieving universality, but also presents big challenges. This paper raises some critical questions about the concepts of "digital libraries" and 'archiving knowledge". It uses a basic life-cycle approach to digital libraries and considers digital library functions within the cycle of the creation, dissemination, disposal and use of born-digital and digitised content. Different types of digital libraries are identified and challenges in selection, acquisition, organisation, preservation, resource discovery and access are discussed. Technological factors are not the main issue to be addressed. Rather, it is emphasised that political and economic challenges require attention. A rational and holistic discipline of digital resources management is needed to ensure that digital content can be handed down to posterity.

Introduction

Librarianship is a profession of modest people. Not many of us become wealthy or powerful. It is fair to say that few of us chose this profession because we consciously sought wealth or power. And yet beneath that modest demeanour is concealed an immense power: the power that derives from control over the immeasurable wealth of human knowledge. And sometimes it seems that the modest demeanour of the librarian conceals an obsession of almost megalomaniac proportions: the obsession to create a universal library.

Over millennia librarians have striven for universality: complete control of all recorded knowledge, if not through ownership then through bibliographic organisation and systems for universal availability and access. In case this appears exaggerated, here are some examples:

- The Library of Alexandria, founded in the 3rd century BC by Ptolemy II of Egypt. By royal decree all books entering the city were confiscated and deposited in the Library. There they were copied, and the copies were returned to the owners. The Ptolemies wanted a universal library (Wikipedia, The Free Encyclopedia, 2008c).
- The bibliography entitled the *Bibliotheca universalis* complied by the Swiss scholar Conrad Gessner. This work was intended to be a catalogue of all writers who had ever lived, with the titles of their works (Wikipedia, The Free Encyclopedia, 2008b). Gessner is known today as the "Father of Bibliography".
- The great national libraries that emerged with ambitions of universality in the 19th century, for example the Department of Printed Books of the British Museum in London, the Bibliothèque Nationale in Paris, and the Library of Congress in Washington DC. The latter claims to be the world's largest library. Currently it holds over 138 million items (including 32 million books) on over 1000 km of shelving (Library of Congress, 2008).
- The *Répertoire bibliographique universel*, founded in 1895 by the Belgian internationalists Paul Otlet and Henri la Fontaine. It was intended to be a universal bibliography of all recorded knowledge. In 1919, when it moved to special quarters in a building in Brussels called the Mundaneum, they had created a database of 12 million catalogue cards (Wikipedia, The Free Encyclopedia, 2008d).
- Theodore A Besterman's *World bibliography of bibliographies,* first published in 1939-40 (Besterman, 1939/40).
- The IFLA (International Federation of Library Associations and Institutions) programmes of Universal Bibliographic Control (UBC) (Anderson, 1974), established in 1974, and Universal Availability of Publications (UAP), established in 1982 (Oakeshot & White, 1984).
- Google, with its mission to "organise the world's knowledge and make it universally accessible and useful" (Google, 2008); the addition of the Google Print and Google Scholar facilities tells us that Google takes this ambition seriously.
- Various comprehensive international digital libraries, including a project partially funded by the European Commission to create a European Digital Library (Leyden, 2006) and the World Digital Library proposal by the Library of Congress

1. Peter Lor, PhD, is Secretary General, International Federation of Library Associations and Institutions, the Hague,The Netherlands, and Extraordinary Professor, University of Pretoria, South Africa

(Library of Congress, 2006a).

Admittedly not all of these examples were initiated by librarians, but they undoubtedly embody the librarian's ideal of universality. Some common themes run through them:

- They reflect a deep-seated human need to hold fast to concrete things in the flux of time.
- Idealism and obsession go hand in hand: everything, for ever, for all.
- Obstacles both financial and logistical are encountered which impede the accumulation of all the desired material. The universal collector simultaneously needs to look back in time (to complete the collection retrospectively), and forward (to keep up with the world's ever-growing output of recorded knowledge).
- Ever-present dangers threaten the survival of the collections or systems.

From this quite limited set of examples we can also trace a rough development pattern:

- First, attempts to collect and preserve everything (*ownership)*.
- Then, universal collections not being feasible, there is a scaling down of the ambition. Attention shifts to attempts to record everything (*bibliographic control*).
- But since a bibliographic reference is no substitute for the document it describes, there follow cooperative attempts to create mechanisms for universal availability of materials held in multiple collections (*access*).

At the risk of some over-simplification it is suggested that each time a new technology for the recording and dissemination of knowledge arises, the cycle starts afresh. Thus the invention of printing led to fresh attempts to create universal collections – which proved not feasible. Hence the attention shifted first to bibliographic control and then to resource sharing for availability. Now digital technology reawakens the dream of the universal collection, as we see in Google's massive digitisation initiatives and those of the European national libraries and the Library of Congress. But is the universal collection a dream, an illusion, or a nightmare?

Against this background the intention in this paper is to raise some critical questions about the concepts embodied in this theme of "digital libraries and archiving knowledge". This is not a technical paper. It presents a basic life-cycle approach to digital libraries, written from the perspective of the political economy of information.

A life-cycle approach to digital libraries and the archiving of knowledge

The emphasis in this paper is on these environmental factors. In the last part of this paper an old-fashioned life-cycle approach is applied to provide a framework for a walk-through of various aspects of the digital library (Fig. 1). In the course of the walk-through some questions are raised and some critical comments are made.

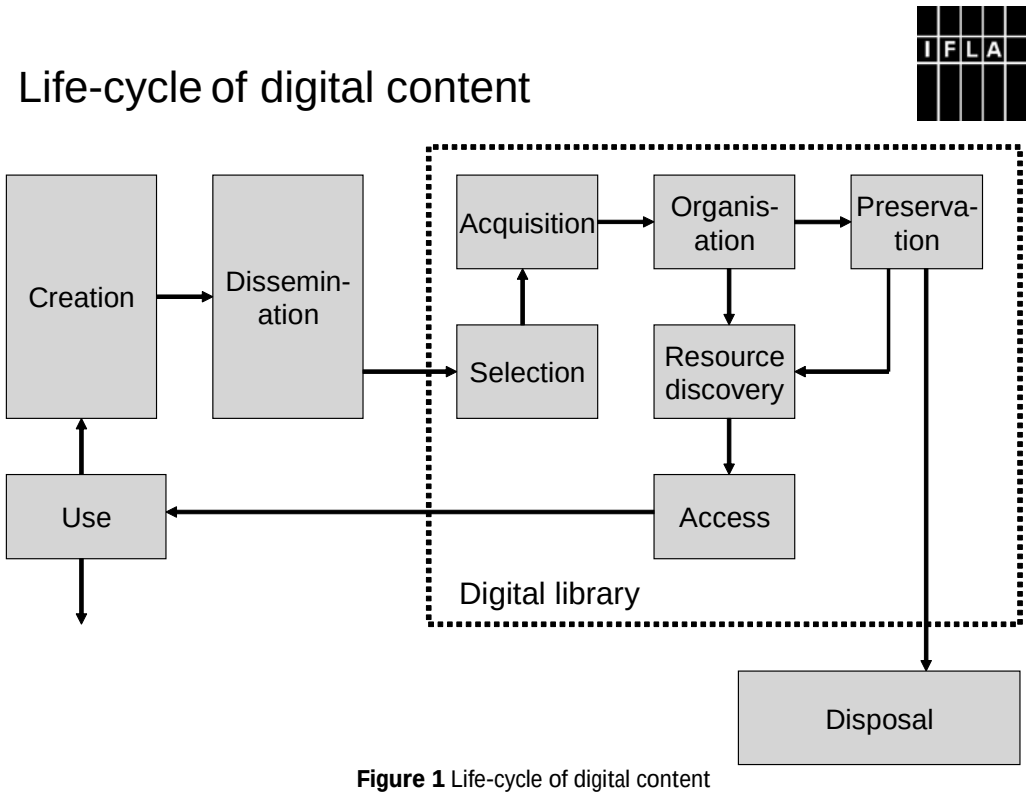

Figure 1 Life-cycle of digital content

Creation

The first question is: who are the creators of knowledge? Is it only those engaged in science and scholarship conducted in accordance with modern (largely Western) paradigms in formal institutional settings? Or is the archiving of knowledge more inclusive, also to cover, for example:

- Scientific and scholarly knowledge derived from non-Western science
- Indigenous or traditional knowledge
- Non-fiction (popular science)
- Pseudoscience
- Music, dance, theatre, fine arts, and crafts
- Literature/fiction
- Popular media content
- Data

Here the inclusive approach is favoured, and the less judgmental term "content" is used. If digital libraries are to reflect the wealth of human knowledge, they should be open to different concepts of knowledge. The boundaries between the categories (all of which can serve as research data) are difficult to draw and will shift over time.

Libraries can play a significant role in the preservation of indigenous knowledge, aiding in its discovery and recording, organising it for use, preserving it, and promoting its appreciation (including respect for the dignity of the communities that produce it) and use (Lor, 2004). The knowledge management skills of librarians should also be put to work in the management of data and other forms of content.

Much has been written about the distinctions between knowledge, information and data. The International Council for science (ICSU) distinguishes between data and information. ICSU's definition of "data" is of interest because it suggests that it encompasses many forms:

> "Data" includes, at a minimum, digital observations, scientific monitoring, data from sensors, metadata, model output and scenarios, qualitative or observed behavioral data, visualizations, and statistical data collected for administrative or commercial purposes" (International Council for Science, 2004: 14).

ICSU's definition of "information" is interesting because it relates information to data:

> "Information" generally refers to conclusions obtained from analysis of data and the results of research. But the distinction between them is flexible and will vary according to the situation. Increasingly the output of research (traditionally viewed as "information") includes data and has become input for other research, rendering the output-input distinction between data and information meaningless" (International Council for Science, 2004: 14).

Modern science is increasingly dependent on very large collections of data, often collected over long periods of time, across national borders, and at great cost by the public sector (e.g. demographic, economic and meteorological data) or the private sector (e.g. Earth observation by remote sensing from satellites, and genome sequencing) (International Council for Science, 2004: 16-17). Such data is typically used in multiple projects. In the past the raw data collected by researchers remained in filing cabinets or on computer storage media in their offices. Today it is becoming more common for raw research data to be made available electronically, for example through a clickable link from an electronic journal article – communication *of* data as distinct from communication *about* data. This makes it possible for data to be re-used by other researchers, to replicate research, verify findings or answer other questions, or for educational purposes. Such re-use, or data mining, requires professional management of scientific data. In this context the concept of "digital curation" has emerged. In the United Kingdom a Digital Curation Centre has been established, which has defined digital curation as follows:

> Digital curation is maintaining and adding value to a trusted body of digital information for current and future use; specifically, we mean the active management and appraisal of data over the life-cycle of scholarly and scientific materials (Digital Curation Centre, 2007).

In the mean time the concept has been expanded to include all forms of digital content and most aspects of digital repositories.

The re-use of research data also raises ethical issues, for example issues of privacy when data from opinion surveys and clinical trials are re-used. Questions of intellectual property also arise: whose data is it and what happens to the copyright if a dataset is submitted for publication along with the schoarly paper? This question was recently addressed in a joint statement of the Association of Learned and Professional Society Publishers (ALPSP) and the International Association of Scientific, Technical and Medical Publishers (STM), which recommended that raw research data, but not processed outputs such as databases, should generally be made freely available (Association of Learned and Professional Society Publishers & International Association of Scientific, Technical and Medical Publishers, 2006).

Dissemination

Content today is disseminated in two forms: analogue and digital. There is so much emphasis on digital media that the importance of analogue materials (ranging from manuscripts through print to a range of analogue audio-visual media) is sometimes overlooked. Libraries will need to maintain extensive analogue collections for the foreseeable future, and provide for voluminous additions to their analogue holdings. This is particularly true of research and national libraries, which have a long-term preservation responsibility. It is no coincidence that major national libraries in highly developed countries (such as the Koninklijke Bibliotheek in the Netherlands) are extending their conventional storage space very significantly.

Digital content can be divided into two categories:

- Digitised content: analogue content that was subsequently digitised: "Digitised [implies] the transformation of the information of the original physical, analogue carrier into a digital form" (Verheul, 2006: 21).
- Born-digital content, which originated and is disseminated digitally: "Born-digital refers to materials which are not intended to have an analogue equivalent of the object, either as the originating source or as a result of conversion to analogue form" (Verheul, 2006: 21).

Digitised content

First, digitised content. Why digitise analogue materials? The key reason is access. There are two dimensions to this: what one may call aggregated access and enhanced access.

- Aggregated access refers to digitisation of very large collections of materials that would otherwise require users to spend much time travelling and searching in numerous antiquarian bookshops and libraries. These very large collections provide access to older or more obscure research literature such as journal back runs that are not held in the researcher's own institution (e.g. JSTOR). They also provide raw materials for social science and historical research, for examples in digitised collections of newspapers,government documents, 18th century popular novels, 19th century street directories, theatre posters and colonial postcards. They add much value by bringing together these materials. They aggregate supply and demand, making it more likely that some user might discover useful content in the "long tail" (Dempsey, 2006) of little-used material.
- Enhanced access refers to digitisation of fragile or vulnerable analogue materials that for reasons of preservation have to be used *in situ* and sparingly. Here value is added by simultaneously reducing pressure of use and providing the user with an enhanced experience, for example the ability to enlarge images, turn pages, and reveal palimpsests.

Digitisation is often suggested as an answer to the problems of preserving analogue material. It can be argued that digitisation is not as such a means of preservation. It is a powerful tool for promoting awareness and appreciation. It enables us to provide access across barriers of time and space. But the digitised medium is vulnerable and ephemeral; if long-term preservation is the objective, preservation of originals and preservation microfilming may be more appropriate.

Digitisation appears to be an unalloyed blessing. But there are some reservations. For example, a considerable number of projects are being undertaken to digitise African heritage material (Britz, & Lor, 2004; Tsebe, 2005). At face value this is a wonderful way of promoting an awareness and appreciation of Africa's rich cultural heritage, but caution is called for. We need to ask critical questions, for example on the ownership of the digitised content, who benefits from the project, and whether the people whose heritage it is will be able to gain access to the digitised content. The ethical considerations should not be overlooked (Lor & Britz, 2004). Similarly, we need to look closely at the intellectual property implications of any digitisation project. Permission is needed from copyright holders before their works are digitised. For this reason most large-scale projects concentrate on material that is in the public domain. But even when the material is in the public domain, it is possible to infringe the moral rights of authors (Oppenheim, 1996). A further question is whether the digitisation project could result in materials that were in the public domain being copyrighted.

Content being digitised is not limited to print originals. Images, sound recordings and video also form part of the recorded heritage of humanity.

Born-digital content

Some born-digital content may have an analogue near-equivalent version that is disseminated simultaneously, for example in the case of electronic newspapers and newsletters, but they are rarely true equivalents. As web publishing becomes more and more sophisticated, the gap between the analogue and digital versions widens because the analogue versions lack the interactive features such as pop-ups, banners, databases and RSS feeds.

The amount of born-digital material is staggering, and it keeps growing. A recent survey by ALPSP of major international academic journal publishers (both non-profit and commercial) found that 90% of their journals are now online, compared with 75% in 2003, while the availability of back issues online has also increased significantly (Cox & Cox, 2006). The migration of scholarly journals to the web is not a simple matter of transferring from an older to a new technology. New business models also have to be found. Furthermore, the advent of the new digital technology and the

possibilities it offers authors and institutions to disseminate their content directly to users, have both opened a Pandora's box of dissatisfaction with the conventional journal publishing system, and provided the means potentially to circumvent it.

Journal publishers today stand accused of presenting serious obstacles to the transmission of content to users, particularly users in developing countries. For example:

- Steeply rising, unaffordable prices
- Unfair licensing schemes
- Double dipping (the client is made to pay twice, first as creator, then as user)
- Excessive profits
- Predatory intellectual property tactics (Britz & Lor, 2003; Lor & Britz, 2005)

Libraries and users in developing countries are most severely affected by these conditions, but even the wealthiest research libraries in the developed countries are affected. It is not surprising that alternatives are being developed, and their multiplication is made possible by rapid developments in information technology. A combination of new technology, outdated business models and greed threatens the survival of the current for-profit journal publishing industry.

Various alternatives are being discussed and explored. These include:

- Not-for-profit aggregators; examples are eIFL (electronic Information For Libraries), J-STOR (JSTOR, the Scholarly Journal Archive) and PERI (Programme for the Enhancement of Research Information, a programme of the International Network for Availability of Scientific Publications, INASP)
- Open access journals
- Institutional repositories
- Discipline or problem-oriented repositories

The open access movement has attracted much attention and wide support from many quarters, including governments, grant-making bodies, and professional organisations. IFLA stated its position on open access in 2003, in its *IFLA Statement on Open Access to Scholarly Literature and Research Documentation* (IFLA, 2003). The statement affirms the importance of comprehensive open access to scholarly literature and research documentation, but without expressing a preference for any particular model.

The business model of commercial and learned society journal publishers may be failing, but open access journals need viable business models too. These will take time to crystallise out. In the mean time, uncertainty will continue.

Scholarly journals are not the only born-digital content of relevance to digital libraries. There is a growing range of other material, including digital broadcast video, business and legal records, social science datasets, physical and biological science data, geospatial data, and websites. The importance of websites as sources of raw research data for historians, political scientists, sociologists, media scientists and other students of the social sciences and humanities is increasing rapidly as the web takes over more and more of the communication functions of printed media such as newspapers and directories. However, the websites are far more ephemeral than the printed sources they replace (*Political communications web archiving…* 2004). Worldwide a vast amount of this material is disappearing into cyberspace on a daily basis. There are many categories of websites, ranging from the personal websites of individuals to institutional, corporate and government websites. Significant recent additions have been blogs and wikis. Blogs range from personal diaries and musings of entirely forgettable individuals to those of scholars, business leaders, activists and prominent politicians (Wikipedia, The Free Encyclopedia, 2008a). By the end of 2004 there were between four and ten million blogs (McGann, 2004), but on 5 June 2008 Technorati (2008) claimed to be tracking 112.8 million blogs. Wikis open up a new mode of collaborative, networked intellectual activity. The most prominent example today is the Wikipedia, with more than 9,7 million articles (2,3 million in English) in 256 languages on 1 March 2008 (Wikipedia, The Free Encyclopedia, 2008e).

The digital library

At this point in the life-cycle of digital content, we come to the role of the digital library proper. This section deals briefly, non-technically and quite selectively with some aspects related to the library functions of Selection, Acquisition, Organisation, Preservation, Retrieval and Access. For the purposes of this discussion it is suggested that roughly speaking there are five kinds of digital library collections:

- Virtual library: A library that operates, or attempts to operate, almost exclusively digitally and to hold only digital collections. Such libraries typically serve a defined clientele (such as the staff of a company) in a delimited field in science, technology or business in which information has a high rate of obsolescence and in which a very high proportion of all information is available digitally, for example the Walden University Library of Indiana University, which has no print collection but only online databases (Barsun, 2005).
- Hybrid library: A high percentage of all academic and research libraries hold both digital and analogue collections. In

many cases, periodical collections are overwhelmingly digital while electronic reserves are growing, but substantial analogue collections remain. In hybrid libraries conscious efforts are made to integrate the analogue and digital components so that users can gain seamless access.

- Virtual heritage library: These libraries digitise materials relating to a particular country, culture or language group, etc. to build very extensive digital collections, generally aimed at pupils, students and the general public. An example is the American Memory, which forms part of the National Digital Library programme of the Library of Congress (Library of Congress, 2006b).
- National digital library: This is typically a component of a country's national library that attempts to collect and preserve the country's born-digital heritage in all fields (i.e. not only patriotica) comprehensively. An example is the "e-Depot" of the Royal Library, the national library of the Netherlands (Oltmans, & Lemmers, 2006).
- Universal digital library: a library that strives for comprehensive coverage across national borders by means of very large-scale digitisation. Examples are the Google Scholar system (Noruzi, 2005), the Library of Congress's World Digital Library (Library of Congress, 2006a) and the European Digital Library (Texier, 2006).

Selection

The sheer volume of digital material and material to be digitised makes selection necessary. Major factors affecting selection decisions are:

- The mission of the institution, for example in a national library there may be a desire to collect the country's digital output comprehensively for preservation, whereas an academic or special library might be much more demand-driven and retain materials only for as long as they are needed by clients
- The anticipated needs of the library users
- Financial resources
- Copyright status: heritage digitisation projects often start with material that is in the public domain, since obtaining permission to digitise material that is in copyright is extremely labour-intensive and uncertain. A large slice of the analogue production of out-of-print material consists of so-called orphan works, of which the copyright owner is unknown or not findable (Libraries and Archives Copyright Alliance, 2007). This situation is made worse by the extension of the term of copyright in the USA, by the Sonny Bono Term Extension Act (Lipinski & Buchanan, 2006), in Europe and in other countries under US influence.
- National or language bias: the digitising libraries or agencies tend to concentrate on materials from their countries and in their languages.

The Google Scholar project illustrates the last three points in particular. Business analysts wonder whether Google will be able to sustain the costs of this project. Publishers are concerned about the digitisation of copyrighted works, even though Google undertakes to make only snippets available online. Librarians generally welcome any scheme that makes a large volume of material readily available, but there are some concerns about excessive dependence on a corporation that may be all too ready to exercise or permit censorship if this is the price to pay for entry into a lucrative market (Lanchester, 2006). Outside the Anglo-American sphere some library leaders, such as the former chief executive of the Bibliothèque nationale de France, Jean-Noël Jeanneney, have expressed concern about the dominance of English-language content and about the monopolistic situation that may arise from Google's position as the provider of both the search engine and the content. These concerns have led to support from the European Commission for a European project, the European Digital Library, as a counterpoise to Google.

Selection implies making choices to avoid being overwhelmed by the sheer volume of material, such as websites or broadcast material. Material may be selected on the basis of our current understanding of what is significant and may be needed in future. Another approach is to do random sampling, to ensure that a representative sample of material from is preserved. For example if subject to legal deposit, the radio and television programmes broadcast in a country could be downloaded on a number of randomly chosen days each year. Thus future media or cultural historians would have a sample of each year's programming to study (Rugaas, 1998).

Acquisition

Acquisition is interlinked with Selection, particularly in digitisation projects, and in many cases with Creation and Dissemination as well.

In the case of analogue content, acquisition usually implied purchase of a physical copy, but acquisition of digital content often means acquiring a licence to access material which remains resident on the server of the publisher or content aggregator. This is in line with the "just in time" service philosophy and the idea that access is more important than ownership. This is now common wisdom among librarians. But there are dangers to the "access over ownership" approach. If a library buys access to an electronic journal it may find that in terms of the licence conditions it has no access to back runs if the subscription is cancelled. Although major publishers no longer impose this sort of condition, many

smaller ones still do (Cox & Cox, 2006). In any case publishers, even very big ones, do not survive for ever. What happens to the back files if an electronic publisher disappears?

Legal deposit can serve as a mechanism for acquisition of digital content by national libraries and other legal depositories. A number of countries have amended their legal deposit legislation to cover online content such as electronic journals and websites. However, the legislation has not been easy to implement. In 1998 South Africa enacted legal deposit legislation that provides for the downloading and archiving of South African websites, but this provision has not yet been put into effect due to financial constraints (Lor, Britz & Watermeyer, 2006). In countries where more progress has been made, various problems have been encountered. Some examples (Conference of European National Librarians and Federation of European Publishers, 2005):

- Legal deposit normally applies only to publications of the country where the deposit is to be made, but in the case of electronic publications, it is often not clear where the publisher is located.
- In legal deposit of print material each issue or edition of a work is normally to be deposited, but what is to be done in the case of continuously updated online works?
- In legal deposit of print material a copy has to be delivered at the expense of the publisher. However, the deposit of an electronic publication may require a costly reformatting exercise, or very frequent downloads. Who is to foot the bill for this?
- The downloading of legal deposit material from the web implies the making of a copy, and to do this may require copyright legislation to be amended as well.

Other legal and ethical problems may arise. The importance of websites as sources of raw research data for historians, political scientists, sociologists, media scientists and other students of the social sciences and humanities is increasing rapidly as the web takes over more and more of the communication functions of printed media such as newspapers, posters, fliers, directories and magazines. However, the websites are far more ephemeral than the printed sources they replace. In some developed countries attempts are being made to archive websites systematically, but worldwide a vast amount of this material is disappearing into cyberspace on a daily basis. A 2004 study suggested that the Internet Archive is not necessarily a reliable and comprehensive repository of websites from developing countries (Thelwall & Vaughan, 2004).

In 2004 the Andrew W. Mellon Foundation funded the Political Communications Web Archiving Project, undertaken in the United States under the aegis of the Center for Research Libraries, Chicago. This project studied the capture, long-term preservation and accessibility of web sites of political groups in various parts of the world. Because they contain information on political movements and conditions in the countries concerned – information that is not readily available elsewhere – such web sites are of interest to political scientists and other social scientists working in area studies. However, these sites are mostly short-lived. They may be closed down by repressive governments or simply disappear after an election because the immediate motivation for operating them has fallen away. Within the project a curatorial group addressed questions relating to copyright and moral issues relating to the harvesting of web sites:

- Should the harvesting institution ask permission before harvesting a site? In many cases the site will have disappeared from the web before an answer is received.
- Could the site owner sue the harvesting institution for violation of its copyright if the site is harvested without permission?
- Should the site be harvested first and put into a "dark archive" pending permission from the site owner?
- And could it be retained in the dark archive if permission is not granted?

Even if there is no legal challenge, ethical problems remain. Here is a hypothetical example: An election is held in country Z. To pacify donor agencies opposition parties are allowed to participate. A few opposition web sites appear during the campaigning period. Predictably, the governing party wins the heavily rigged poll and immediately clamps down on all opposition. The opposition leaders go underground and their web sites disappear. Afterwards internal security agents from country Z visit the harvested web sites to identify and track down opposition politicians. They use the archived content as evidence in the ensuing treason trial. This is hypothetical but not unrealistic.

Other questions studied in the Political Communications Web Archiving Project included the selection of sites, timing of capture (for example, how often to capture the same site), what is an "acceptable level of loss", how to keep different versions, and the drafting of a collection policy and a risk management strategy. The problem of the "deep web" (e.g. password protected content and interactive databases "behind" the surface web, which are much more difficult if not impossible to capture) featured prominently in the discussion, as did the issue of metadata (in which languages, how much human input would be needed) and keyword indexing (*Political communications web archiving*… 2004).

Organisation

This section is limited to some general remarks on organisation for retrieval.

The content of the digital library has to be organised for retrieval, or as it referred to in the context of digital resources, for resource discovery. The term "resource discovery" has merit in that it approaches the problem from the user's point of view: users cannot *retrieve* content if they are not aware of its existence, but they can be helped to *discover* it. It does not matter to users where the content is held; whether in a book, on a CD, or on a server in their own institute, in the library or on the other side of the world. Users whose expectations have been formed by their experience of search engines such as Google expect seamless and instant access, meaning not just bibliographic references (or a peek at the content accompanied by a demand for payment) but delivery to his/her workstation. This means seamless integration not only of disparate bibliographic control mechanisms but also financial and user authentication systems.

It follows from the nature of digital content, much of which is online, that the organisation of digital content cannot be limited to the silos of individual institutions. This would be a waste of the potential of such facilities as institutional repositories. These repositories are there to give greater, extra-institutional exposure to their content. Linking of institutional repositories can take many forms. In the Netherlands, the SURF Foundation, a cooperative body which supports ICT and network services to Dutch higher education and research institutions, set up the Digital Academic Repositories (DARE) programme, a joint initiative of the Dutch universities to make all their research results digitally accessible. The KB (National Library of the Netherlands), the KNAW (Royal Netherlands Academy of Arts and Sciences) and the NWO (Netherlands Organisation for Scientific Research) are also cooperating in this project. It is the first network to link all the institutional repositories of all the universities in one country. The project resulted in a European successor under the name DRIVER.which aims to create a single, large-scale virtual content resource that provides access to all European research materials (SURF Foundation, 2006).

SURFNet manager Leo Waaijers believes that libraries have an important role to play in respect of institutional repositories, for example taking on their management and maintenance, and in the process evolving from libraries to 'libratories' (Waaijers, 2005).

Resource discovery

This means the utilisation of the systems for resource discovery. Although there are some who believe that modern ICTs will lead to disintermediation, which would lead among other results to the elimination of librarians, this is unlikely to happen. Modern ICTs instantly deliver enormous amounts of information to the workstation of the users, but at the risk of overwhelming them. A search on Google for "digital libraries" yielded approximately seven million hits in 0,04 seconds. "Disintermediation" yielded 261.000 hits in 0,24 seconds; "resource discovery" 1,17 million in 0,15 seconds. This is surreal. Of course, in these totals there is an enormous amount of duplication, and there are many false hits. Also, a simple search on these terms in Google is not the best or only way to search for literature – but someone has to teach the users this. For many students, even university staff, if something cannot be "googled" it does not exist.

The critical competence of the client faced with the abundance and diversity of digital resources is information literacy. In a statement entitled *Beacons of the Information Society: the Alexandria proclamation on information literacy and lifelong learning* (IFLA, 2005), an expert group convened by IFLA and UNESCO defined information literacy as follows: Information literacy –

> comprises the competencies to recognize information needs and to locate, evaluate, apply and create information within cultural and social contexts;
>
> is crucial to the competitive advantage of individuals, enterprises (especially small and medium enterprises), regions and nations;
>
> provides the key to effective access, use and creation of content to support economic development, education, health and human services, and all other aspects of contemporary societies, and thereby provides the vital foundation for fulfilling the goals of the Millennium Declaration and the World Summit on the Information Society; and
>
> extends beyond current technologies to encompass learning, critical thinking and interpretative skills across professional boundaries and empowers individuals and communities

Librarians are ideally placed to provide information literacy education, including education about copyright. (Copyright education on university campuses is best not left to publishers' representatives or reproduction rights organisations, eager as they may be to offer it free of charge, since they commonly forget to mention details such as "fair use".) In addition, librarians provide awareness and alerting services and last but not least, user support, which includes motivating and awareness-raising in the user community and educating and counselling individual users.

Preservation

How long should resources be preserved? The national library of a country with a relatively short library history can serve as an example. The print collections of the National Library of South Africa, which is young by European standards, go back more than 250 years. This implies a time horizon of 250 years, at least conceptually, for the preservation of contemporary material. In principle there is no reason why electronic media collected in terms of legal deposit legislation should be preserved for shorter periods than print, but a time horizon of 250 years must be almost unimaginable to IT professionals, who may be more inclined to think of a decade or two.

The long-term preservation of digital content presents formidable technological challenges. Not only does the content have to be preserved (including in principle, text, images, embedded files, hyperlinks, the deep web, and password-protected content), but also the context (e.g. the "look and feel" of the pages and pop-up advertising). Clearly, magnetic or optical media are more suitable for preserving content and context than print or microfilm, but at this stage of development these media do not enjoy life expectancies that remotely approach the 250 year time horizon (Dollar, 2000; Lor, Britz & Watermeyer, 2006).

The physical deterioration of storage media can be countered by refreshing stored files (copying them onto fresh media). More serious problems are caused by the obsolescence of the physical hardware, and associated low-level software (micro code), on which the information is stored, and the obsolescence of the software and applications used to present, or display, the information (Lor, Britz & Watermeyer, 2006). These issues, which are complicated by the absence of standards for long-term archiving of electronic material, and various strategies such as bit-level preservation, conversion, migration and emulation, are dealt with extensively in a growing body of literature (Verheul, 2006: 51-55).

But this is not merely a question of technology and standards, important as they are. It is also a question of management and of coordinated national and international strategies. The essential problem has been well formulated in ICSU's report, *Scientific data and information* (International Council for Science, 2004).

> 'Physical' libraries have historically been considered as the repositories and long-term guardians of scientific publications. Commercial publishers themselves have also accepted responsibility for archiving publications, although in a competitive commercial world, publishing houses cannot guarantee long-term continuity. The complementarity between the short to medium term role of publishers and the longer-term function of libraries has been a reasonable guarantee of the preservation of the scientific record to date. However, with the advent of new electronic publishing paradigms and the consequent upheaval in scientific information exchange, it is now less clear where the ultimate responsibility for the archiving of scientific information will lie. Many centralized and decentralized models are being developed but their longterm viability is difficult to assess. In the meantime, there is a crucial role for traditional libraries to play in ensuring continuity in relation to electronic scientific publications. Libraries also have a key role to play in setting up and managing institutional repositories that can organize and preserve institutional research output in digital form.

Much attention is currently being paid to options for the long-term preservation of digital scholarly content, e.g. legal deposit, "trusted digital repositories" (Hank, 2006; Research Libraries Group, 2002) and the LOCKSS (Lots of copies keep stuff safe) concept (LOCKSS 2008). At the national level, fully operational national digital repositories currently exist in a number of national libraries, e.g. Australia, the Netherlands and the United Kingdom. But how many countries will be able to afford to set up and operate their own national repositories? It is fair to say that many developing countries will not be able to do so, at least with the expertise and technology currently available. Many national libraries form part of a national framework, such as the National Digital Information Infrastructure and Preservation Program (NDIIPP) in the United States and e-Helvetica in Switzerland. Regionally and internationally libraries are working together in consortia such as the International Internet Preservation Consortium (IIPC). On a smaller scale the IFLA CDNL Alliance for Bibliographic Standards (ICABS) groups together a number of leading national libraries in the USA and Western Europe (Verheul, 2006: 56-64). But all this activity is largely confined to the wealthy countries.

Disposal

The life-cycle diagram (Figure 1) has a "Disposal" component at bottom left. We cannot collect everything and keep it for ever. Some discarding or "deselection" inevitably has to take place. This should be done in accordance with pre-determined policies and schedules, not as a response to crisis situations.

Access and use

Much of what has to be said about access has already been dealt with under resource discovery. In the case of analogue material, the user usually has to follow a three-step process to gain access to useful content: first a search in catalogues and bibliographical tools to identify relevant material, secondly to determine where it is located, and thirdly to find it on

the shelf or obtain a photocopy. This can be a recursive process, for example if it is necessary to order an interlibrary loan. In the case of digital content these steps can be carried out in a series of clicks – at least in theory. Intellectual property rights and economic barriers may intervene and the user may find words such as the following on his/her screen:

> **This item requires a subscription to the *Journal of ...***

An exploration of the implications and ramifications of this barrier, which would require a discussion of licensing of digital content, big deals, access to knowledge, free trade agreements, the information commons, open access models, the Budapest Declaration, Creative Commons, copyleft, etc., is beyond the scope of this paper.

A further barrier that may be encountered is that of censorship. Although it may seem self-evident that freedom of access to information and freedom of expression are essential for the development of a well-educated, information-literate population that is able to participate actively in the knowledge society, there are countries that aspire to develop as knowledge societies but severely restrict freedom of expression, particularly on the Internet. It has been argued that these countries may conceivably make progress towards the information society, but that the knowledge society proper is beyond their reach. A knowledge society requires a high degree of creativity, intellectual curiosity, openness to divergent views and critical interaction, which depend on intellectual freedom (Lor & Britz, 2007). Creation is dependent on use. However, threats to the intellectual freedom can arise at any decision-making point in the digital content chain, from Creation through Dissemination and the various curation activities to access and use. IFLA in 2002 issued an Internet manifesto (IFLA, 2002) stating that access to the Internet and all its resources should be consistent with the Universal Declaration of Human Rights, particularly article 19:

> Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

This right should not be taken for granted in digital libraries, but requires constant vigilance.

Conclusion: are we approaching the digital dark ages?

It has been said that our grandchildren will be the first generation in the modern era that will be unable to pass on a pictorial record of us, their grandparents, because the colour photographs that we have all been making during the past few decades will all fade away. This is one manifestation of a law of technology which states that the more sophisticated a technology is, the more catastrophic the consequences when it fails.

Will this fate also befall the billions of photographs being taken today by the ubiquitous digital camera? Does this threat also loom over our digital heritage? Are we heading for what has been called the "digital dark ages" (Kuny, 1997)? It has become a popular topic, highlighted in an issue of *Newsweek,* which cited the well-known near-loss of 1972 Landsat images, as well as mind-boggling statistics on the amount of information produced per year: five million exabytes of data – enough to fill 37,000 buildings the size of the U.S. Library of Congress[2] (White & Hastings, 2006). The emphasis, as always in these stories, is on the volume of information currently produced – the threat of information overload – and the fragility of their physical carriers – the threat of destruction.

But the European dark ages from which this metaphor is derived were not characterised by a huge output of information and they were not triggered by a massive technological shift. In the dark ages, things simply fell apart politically and economically. Barbarians moved in, Rome was sacked, large cities were depopulated by economic factors as well as by wars and disease, and culture and learning retreated into monasteries and other isolated centres.

Similarly, it would be a mistake to see technological factors as the main issue to be addressed. Reporting on the National Digital Information Infrastructure and Preservation Program (NDIIPP), Abby Smith wrote:

> The past five years have shown that the "real challenges" in digital preservation are not primarily technical or procedural: they are the policies, the politics, and the economic drivers of digital preservation that serve to

2. This figure is taken from a report by the School of Information Management and Systems of the University of California at Berkeley: The report states: "Print, film, magnetic, and optical storage media produced about 5 exabytes of new information in 2002. Ninety-two percent of the new information was stored on magnetic media, mostly in hard disks. [...] If digitized with full formatting, the seventeen million books in the Library of Congress contain about 136 terabytes of information; five exabytes of information is equivalent in size to the information contained in 37,000 new libraries the size of the Library of Congress book collections." (How much information? 2003). An exabyte is 10^{18} bytes (a million terabytes or a million million megabytes).

> divide stakeholders as often as they unite them in a common cause. It is no longer true […] that content producers, distributors, and consumers do not understand the risk of data loss. […] But their interests in preservation at best overlap, or appear to be in conflict, because they do not share common understandings of the value of that information – for whom, for how long, for what purpose (Smith, 2006).

Modern librarianship has learned to live with the tension between short-term funding constraints and long-term goals (Smith, 2006). A time horizon of 250 years, even 100 years, for the preservation of digital content seems impossibly remote. But today we can at least set ourselves a target to develop a rational and holistic discipline of digital resources management that will enable us to hand on a significant and representative proportion of the digital content to the next generation.

To the preservation challenge we should add two specific areas where a common understanding is lacking: intellectual property rights and freedom of expression. Here too there are threats that can give rise to the digital dark ages, and here too the problem is a lack of common understandings. Interests have to be balanced. Reason has to prevail in the cultural, social, economic and political spheres.

Digital libraries in one form or another will play an important role in achieving the "shared vision to bridge the digital divide and create a truly global Information Society" to which Kofi Annan referred in his introduction to the Geneva Declaration of the World Summit on the Information Society (WSIS, 2003, p.2). They will make mass access to knowledge possible for millions that are currently unserved, provide enhanced information services to scholars, students and decision-makers, and give fresh exposure to authors and writings, languages and cultures that are at risk of being buried in the deluge of information. But these benefits come with a price tag. Digital technology is not politically or culturally neutral. The price of a "truly global Information Society" is constant vigilance and patient consensus-building with all stake-holders to ensure that the technology is harnessed judiciously, responsibly and fairly.

Acknowledgement

I am grateful to Retha Claasen-Veldsman for helping me to adapt this article, originally presented as a keynote address given at the InfoVision 2006 Knowledge Summit, Bangalore, 28 and 29 September 2006, for publication.

References

Anderson, D. (1974). *Universal bibliographic control: a long term policy, a plan for action.* Pullach/München: Verlag Dokumentation.

Association of Learned and Professional Society Publishers & International Association of Scientific, Technical and Medical Publishers. (2006). *Databases, data sets, and data accessibility – views and practices of scholarly publishers: a statement by the Association of Learned and Professional Society Publishers (ALPSP) and the International Association of Scientific, Technical and Medical Publishers (STM).* [Online]. Available: *http://www.alpsp.org/ngen_public/default.asp?ID=202* (Accessed 2008-06-07)

Barsun, R. (2005). The Walden University library: reaching out and touching students. *Internet reference services quarterly* 9(1/2): 93-109.

Besterman, T.D.N. (1939/40). *A world bibliography of bibliographies*. 2v. London: the Author.

Britz, J.J. & Lor, P.J. (2003). A moral reflection on the information flow from South to North: an African perspective. *Libri* 53(3): 160-173.

Britz, J.J. & Lor, P.J. (2004). A moral reflection on the digitization of Africa's documentary heritage. *IFLA journal* 30(3): 216-223.

Conference of European National Librarians and Federation of European Publishers. (2005). *Statement on the development and establishment of voluntary deposit schemes for electronic publications*. [Online]. Available *http://www.nlib.ee/cenl/docs/05-11CENLFEP_Draft_Statement050822_02.pdf#search=%22CENL%20FEP%20Voluntary%20deposit%22* (Accessed 2008-06-07)

Cox, J & Cox, L. (2006). *Scholarly publishing practice: academic journal publishers' policies and practices in online publishing*; 2nd survey. Clapham: Association of Learned and Professional Society Publishers. Executive summary. [Online]. Available: *http://www.alpsp.org/ngen_public/article.asp?id=200&did=47&aid=269&st=&oaid=-1* (Accessed 2008-06-07)

Dempsey, L. (2006). Libraries and the long tail. *D-Lib magazine* 12(4). [Online]. Available: *http://www.dlib.org/dlib/april06/dempsey/04dempsey.html* (Accessed 2008-06-07).

Digital Curation Centre. (2007). *About the DCC*. [Online]. Available: *http://www.dcc.ac.uk/about/* (Accessed 2008-06-07).

Dollar, C.M. (2000). *Authentic electronic records: strategies for long-term access*. Chicago: Cohasset Associates.

Google. (2008). *Corporate information: company overview.* [Online]. Available: *http://www.google.com/corporate/* (Accessed 2008-06-07).

Hank, C. (2006). Digital curation and trusted repositories, seeking success: JDCL Workshop report. *D-Lib magazine* 12(7/8). [Online]. Available: *http://www.dlib.org/dlib/july06/hank/07hank.html* (Accessed 2008-06-07).

How much information? (2003). *Executive summary.* [Online]. Available: *http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm#report* (Accessed 2008-06-07).

IFLA. (2002). *The IFLA Internet manifesto*. [Online]. Available: *http://www.ifla.org/III/misc/im-e.htm* (Accessed 2008-06-07).

IFLA. (2003). *IFLA statement on open access to scholarly literature and research documentation*. [Online]. Available: *http://www.ifla.org/V/cdoc/open-access04.html* (Accessed 2008-06-07).

IFLA. (2005). *Beacons of the information society: the Alexandria proclamation on information literacy and lifelong learning*. [Online]. Available: *http://www.ifla.org/III/wsis/BeaconInfSoc.html* (Accessed 2008-06-07).

International Council for Science. (2004). *Scientific data and information: report of the CSPR Assessment Panel*. Paris: ICSU [Online]. Available: *http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf#search=%22CSPR%20ICSU%20Scientific%20Data%22* (Accessed 2008-06-07).

Kuny, T. (1997). A *digital dark ages? Challenges in the preservation of electronic information*. Paper presented at a Workshop during the 63rd IFLA Council and General conference, Copenhagen, August 1997. [Online]. Available: *http://www.ifla.org/IV/ifla63/63kuny1.pdf* (Accessed 2008-06-07).

Lanchester, J. (2006). Big Google is watching you. *Times online*, 29 January 2006. [Online]. Available: *http://www.timesonline.co.uk/article/0,,2092-2014215,00.html* (Accessed 2008-06-07).

Leyden, J. (2006). European Digital library is go. *The Register: Internet and Law*, Friday 3 March 2006. [Online]. Available*: http://www.theregister.com/2006/03/03/european_digital_library_goes_live/* (Accessed 2008-06-07).

Libraries and Archives Copyright Alliance. (2007). Statement on orphan works. [Online]. Availabel: *http://www.cilip.org.uk/NR/rdonlyres/E6F612ED-6CE1-4723-8348-CB7162D983C2/0/LACAorphanworksstatementFINAL19dec07.pdf* (Accessed 2008-06-07).

Library of Congress. (2006a). *Library of Congress launches effort to create world digital library.* [Online]. Available: *http://www.loc.gov/today/pr/2005/05-250.html* (Accessed 2008-06-07).

Library of Congress. (2006b). *About American memory.* [Online]. Available: *http://memory.loc.gov/ammem/about/index.html* (Accessed 2008-06-07).

Library of Congress. 2008abc About the Library. [Online]. Available: *http://www.loc.gov/about/facts.html* (Accessed 2008-06-07).

Lipinski, T.A. & Buchanan, E.A. (2006). The impact of copyright law and other ownership mechanisms on the freedom of enquiry: infringements on the public domain. *Journal of information ethics* 15(1): 47-59.

LOCKSS. (2008). Lots of copies keep stuff safe. [Online]. Available: *http://www.lockss.org/lockss/Home* (Accessed 2008-06-07).

Lor, P.J. (2004). Storehouses of knowledge? The role of libraries in preserving and promoting indigenous knowledge. *Indilinga: African journal of indigenous knowledge systems* 3(1): 45-56.

Lor, P.J. & Britz, J.J. (2004). Digitization of Africa's documentary heritage: aid or exploitation? *Journal of information ethics* (Fall): 78-93.

Lor, P.J. & Britz, J.J. (2005). Knowledge production from an African perspective: international information flows and intellectual property. *International information & library review* 37: 61-76.

Lor, P.J. & Britz, J.J. (2007). Is a knowledge society possible without freedom of information? *Journal of Information Science* 33(4):387-397.

Lor, P.J., Britz, J.J. & Watermeyer, H.C. (2006). Everything, for ever? The preservation of South African websites for future research and scholarship. *Journal of information Science* 32(1): 39-48.

McGann, R. (2004). *The blogosphere by the numbers*. ClickZ stats. [Online]. Available: *http://www.clickz.com/showPage.html?page=3438891* (Accessed 2008-06-07).

Noruzi, A. (2005). Google Scholar: the new generation of citation indexes. *Libri* 55(4): 170-180.

Oakeshot, P & White, B. (1984). *The impact of new technology on the availability of publications: report to the International Federation of Library Associations and Institutions Universal Availability of Publications (UAP) Programme…* Wetherby: IFLA International Programme for UAP, British Library Lending Division.

Oltmans, E. & Lemmers, A. (2006). The e-Depot at the National library of the Netherlands. *Serials* 19(1): 61-67.

Oppenheim, C. (1996). Moral rights and the electronic library. *Ariadne* (4). [Online]. Available: *http://www.ariadne.ac.uk/issue4/copyright/* (Accessed 2008-06-07).

Political communications web archiving: an investigation funded by the Andrew W. Mellon Foundation. Final report. (2004). Chicago: Center for Research Libraries. [Online]. Available: *http://www.crl.edu/content/PolitWebReport.htm* (Accessed 2008-06-07).

Research Libraries Group. (2002). *Trusted digital repositories: attributes and responsibilities*. An RLG-OCLC report. [Online]. Available: *http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf* (Accessed 2008-06-07).

Rugaas, B. (1998). "*The end of all and forever": on the acquisition policies of national libraries and the future of legal deposit material*. Unpublished paper 39-NAT-3-E presented to the 1988 IFLA Conference, Sydney, Australia.

Smith, A. (2006). Distributed preservation in a national context: NDIIPP at midpoint. *D-LIB magazine* 12(6). [Online]. Available: *http://www.dlib.org/dlib/june06/smith/06smith.html* (Accessed 2008-06-07).

SURF Foundation. (2006). SURF DARE project gains European-wide adoption. [Online]. Available: *http://www.surffoundation.nl/smartsite.dws?ch=ENG&id=10935* (Accessed 2008-06-07).

Technorati. 2008. About us. [Online]. Available: *http://www.technorati.com/about/* (Accessed 2008-06-05).

Texier, B. (2006). Construction européenne d'une bibliothèque numérique. *Archimag* (195): 20-23.

Thelwall, M. & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet archive. *Library & information science research* 26(2): 162-176.

Tsebe, J. (2005). *Networking digital heritage: Africa*. Paper 157-E, presented at the IFLA World Library and Information Congress, Oslo, Norway, 14-18 August 2005. [Online]. Available: *http://www.ifla.org/IV/ifla71/papers/157e-Tsebe.pdf* (Accessed 2008-06-07).

Verheul, I. (2006). *Networking for digital preservation: current practice in 15 national libraries*. München: K G Saur.

Waaijers, L. (2005). From libraries to 'libratories'. *First Monday* 10(12). [Online]. Available: *http://www.firstmonday.org/issues/issue10_12/waaijers/index.html* (Accessed 2008-06-07).

White, P. & Hastings, M. (2006). The digital dark age. Newsweek international, June 26, 2006. [Online]. Available: *http://0-find.galegroup.com.innopac.up.ac.za/itx/retrieve.do?contentSet=IAC-Documents&resultListType=RESULT_LIST&qrySerId=Locale%28en%2CUS%2C%29%3AFQE%3D%28JN%2CNone%2C24*

%29%22Newsweek+International%22%3AAnd%3ALQE%3D%28DA%2CNone%2C8%2920060626%24&sgHitCountType=None&inPS=true&sort=DateDescend&searchType=PublicationSearchForm&tabID=T003&prodId=AONE&searchId=R1&currentPosition=15&userGroupName=up_itw&docId=A147222654&docType=IAC (Accessed 2008-06-07).

Wikipedia, The Free Encyclopedia. (2008a). *Blog*. [Online]. Available: *http://en.wikipedia.org/wiki/Blog* (Accessed 2008-06-07).

Wikipedia, The Free Encyclopedia. (2008b). *Conrad Gessner.* [Online]. Available: *http://en.wikipedia.org/wiki/Conrad_Gessner* (Accessed 2008-06-07).

Wikipedia, The Free Encyclopedia. (2008c). *Library of Alexandria*. [Online]. Available: *http://en.wikipedia.org/wiki/Library_of_Alexandria* (2008-06-07).

Wikipedia, The Free Encyclopedia. (2008d). *Paul Otlet*. [Online]. Available: *http://en.wikipedia.org/wiki/Paul_Otlet* (Accessed 2008-06-07).

Wikipedia, The Free Encyclopedia. (2008e). *Wikipedia: multilingual statistics*. [Online]. Available: *http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics* (Accessed 2008-06-05).

WSIS. (2003). World Summit on the Information Society, Geneva 10-12 December 2003. *The Geneva declaration of principles and plan of action.* Geneva: WSIS Executive Secretariat.