Introduction to Digital Libraries – Memex of the Future

Peter Linde

Blekinge Institute of Technology, Sweden

1. Historical Flashbacks and Definitions

1.1. The problem that would not go away

After Johann Gutenberg's fabulous invention, in the middle of the 15th century, libraries, as we know them, slowly started to appear. At first as private collections but in the 17th and 18th century developing into college and university libraries, as well as libraries of scientific societies. Libraries were created to supply library patrons with relevant material, collecting and protecting books and manuscripts. The texts just started to add and add and add...

During the 17th century more then one million titles were printed and in the same century, a number of scientific societies were established in Europe. In the following century they multiplied but also became more specialised. The societies adjusted to how science diverged into new aspects and new fields of academic study, each with its own agenda and methods that not necessarily were comprehensible to specialists in other branches.

It´s during the expansive and experimental 18th century, that voices of complaints are raised more and more often regretting the flood of scientific literature that seemingly have no end. Laymen and scientist alike were drowning in a flood of information that just seemed impossible to handle. And that was not enough! Complaints about difficulties getting hold of primary sources and the low general quality of scientific output were not uncommon either [1]. Does it sound familiar? Since the 18th century complaints of info-overload have been legio. Contemporary scientific society tried to handle the problem by exchanging abstracts, constructing bibliographies and by publishing scholarly journals.

Today the problem is more or less the same. It has not gone away. Instead we have invented new tools and technologies for dealing with the problem. With the coming of computers in the mid 20th century visionary people proposed ideas and solutions for how to solve the info overload dilemma. One of these was Vannevar Bush.

Mr. Bush was Director of the American Governments Office of Scientific Research and Development and had been a central figure in the development of nuclear fission and in the development of producing the first atom bomb. In an article in "The Atlantic Monthly"[2] in July 1945 he addressed scientists in general and physicists in particular to rise to the great task, now when the war effort was over, of perfecting instruments to access and command the cultural and scientific knowledge handed down by generations. As scientists

before him Vannevar Bush was troubled about the growing mountain of research and scientific specialisation which made it impossible for readers to keep track, much less remember and locate necessary facts. In his article Bush envisioned in the not so far future, things like personal terabyte discs and scanning, only he thought it would happen on microfilm; he talked about speech recognition; he foresaw personal computers with great computational power; he foresaw hyperlinks but called it associative indexing; he envisioned the whole scientific self archiving movement and spelled it "The Memex".
The Memex was a device for personal use – a private file system where the individual could store all books, records and communications needed for private and professional life. It was mechanised and could be consulted and trusted to deliver answers with considerable speed. The Memex was built into a desk with slanting translucent screens on top. The screen and a keyboard, plus a set of buttons and levers were the only thing that made it stand apart from any other piece of furniture. Most of the Memex contents, such as books and articles would be available on microfilm and thus inserted into the machine. But the really outstanding feature of the whole gadget was its capability of tying two items together by associative indexing. This new vision of things to come was to replace the old system of indexing where objects can be located by classification systems only. The human mind doesn't work that way, argued Vannevar Bush, it operates by association. It goes from one item instantly to another. The Memex, Bush suggested, would work so that any item may be called at will to select another immediately and automatically.
These ideas have made Vannevar Bush one of the founding fathers of the World Wide Web. He is always mentioned as one of the important figures that inspired the likes of Ted Nelson and Tim Berners-Lee. But his visions are not only applicable to hyperlinks of the web. The Memex machine is a fantastic index finger from the past pointing in the direction of Digital Libraries and a realisation of the most recent attempt to solve the information overload problem!

1.2. Digital Libraries – Definitions and circumscriptions

The term "Digital Library" is used in articles for the first time in the early 1990s. Since then research and practice in Digital Libraries has become standard. Plenty research funding, especially in the United States and United Kingdom, has made sure that both researchers and librarians have been actively involved in Digital Libraries projects. Conferences and journals in the topic are thriving. A search in Google on the exact phrase "Digital Libraries" returns almost 2 million hits in early 2006. But what has happened in these 15 years since the term first was used, that forced this explosive activity? And what does the concept "Digital Library" really mean?
Well, to answer the first question, what happened was that the flow of information got even worse as the Internet established itself as the number one channel for exchanging data. The World Wide Web (WWW) and the HyperText Transfer Protocol (HTTP) became the communication tool of choice and made producing and disseminating data so much easier. In the early 1990s the first web-browser appeared, pioneered by Tim Berners-Lee at CERN. With the web-browser came the capability to use hyperlinks. In the middle of the 1990s Netscape made browsing a possibility for everyone. At the start of the new millennium we have, at least in the developed countries, enough bandwidth and connection possibilities,

for sending and exchanging very heavy loads of data. We also have new and pretty stable standards for structuring and exchanging data such as the Z39.50 communications protocol designed to support searching and retrieval of full-text documents, bibliographic data, images, multimedia etc. in a distributed network environment, plus we have the Open Archive Initiative – Protocol for Metadata Harvesting. We also can use proven standards for metadata handling such as Dublin Core (DC), Encoded Archival Description (EAD), Metadata and Encoding Transmission Standard (METS). We got better authoring tools and software solutions that made life on the web so much easier.
So, to answer shortly, what happened during the last 15 years was that the tools for realising the Memex started to present themselves one by one.
I think Vannevar Busch would have considered the concept of the Digital Library as partly fulfilling his dream of the Memex. But of course there are many interpretations of what a Digital Library really is. There is no definitive definition of the term. It is widely used and there is no certainty that when discussing Digital Libraries two people will mean exactly the same thing. For example the term "Hybrid libraries" is sometimes used to define a library where digital and printed information co-exists. The forms and shapes of the Digital Library is manifold. It can provide access to digital content only but also be a hybrid that delivers non-digital content parallel to digital content [3]. The Digital Library term has during the last decade become some sort of umbrella term for a diverse array of information projects. So there is no way we can exactly say what the term really means but we can try to sort out some sort of pattern of and circumscription around the phenomena.
There have been an abundance of attempts to define the term during its short life time. After consulting literature [4] and reflecting on applied practices, I think it is safe to say that the concept of Digital Libraries can be divided into two main domains – the researchers domain and the librarians domain. The main difference is that while librarians focus on service and sees the Digital Library as an institution, the researchers focus on the content collected on behalf of and served to special user communities. Both these domains have their own definitions of the phenomena. And neither of these definitions care to deal with the abundance of services on the world wide web that identify themselves as Digital Libraries; and I'm referring to everything from booksellers catalogues, library Online Public Access Catalogs (OPACs) to Proprietary electronic databases such as Inspec, ISI, Springer Link etc. These services and others, like the new mega project "Google Library" can partly connect to certain aspects of the two major definitions very well but most often fail to comply with the full definition of either one.

1.3. Research and praxis

The massive interest and upswing for telecommunication and computer studies in the late 1990s and the availability of funds for such hot topics as Digital Libraries studies across scientific fields lay the foundation for a variety of research into the Digital Library topic. Be it applications, protocols and standards, social aspects, user behaviour, preservation studies etc. It is by nature a interdisciplinary topic and because of that there are problems of definition. The main problem with defining it is the second part of the term – "Library". We can pretty much agree on the definition of the term "Digital" as in "using digits […] applied

to a computer which operates on data in the form of digits or similar discrete elements"[5]. It is when we come to the "library" part where all the trouble begins.

Librarians tend to speak for a broader definition of the term "Library". They see a library as an organisation that secures the selection, conservation, organisation, preservation and the access to information that is vital for the members of the specific organisation. Librarians and Libraries carry a long history and tradition that has been somewhat cemented during the centuries. With the coming of the Internet and digital media, for librarians this is only yet another delivery channel for yet another media.

Researchers most often favour a narrower definition of the library concept. For them a library could be any room containing a small or large amount of books or data discs or tape cassettes. Researchers seldom care for the social and institutional context of the term "Library". Their emphasis is tilted towards databases and how to collect, retrieve, organise and access the information.

But through the 1990s definitions of Digital Libraries have broadened in scope even if there is no definitive one. Trying to summarise definitions given by Research initiatives, Science foundations and Digital Library researches, three elements seems to be necessary: 1. There must be some sort of organised collection. 2. It can be partly bibliographical but full-text files of the data, if it is an article or manuscript etc., is now frequently added and required in various formats. 3. The collection is organised for a group or community of users.

Much of Digital Library research is focused on database structure, data mining, retrieval algorithms, filtering and network architecture. That sort of research is based on the assumption that some users need certain enabling technologies to successfully manipulate specific content. The notion of a certain community is problematic because the definition in itself does not contain the criteria for defining what a user community is. Another definition, that is seldom mentioned, is the assumption that Digital Libraries only operate in distributed environments. But then other definitions would include a CD-ROM with digitised books in a certain subject area as an example of a Digital Library [6].

In 1998 Donald J. Waters [7] presented the first, short and workable librarians definition of Digital Libraries, which also was adopted by the Digital Library Federation (DLF) who´s members are major American universities as well as Library of Congress and British Library. It reads:

> *"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."*

There is quite a distinction here compared with the research oriented definition. Here the focus is on the Digital Library as an organisation providing information services in digital form and also taking responsibility for preservation and integrity of the collection. The definition is much broader. But the fact that a library is calling itself for a Digital Library does not usually mean that all its services are digital. It usually means that some parts of the information services are digital. Very few libraries are Digital Libraries in the sense that their services are digital only.

In an article from 1996 Ross Atkinson [8] predicts the new role of academic libraries that is partly realised today. He says that technology will provide libraries with the ability to distribute and make scholarly publications accessible more effectively and that mediation of scholarly information will be taken over from the the large commercial publishers by the academic library community. This is important he argues, because information technology does not promote access. It promotes control. And this control has and will be used by commercial information proprietors to increase revenue as much as possible if not the control is used to promote access.

Academic libraries and organisations can and should assume scientific publishing responsibilities in order to promote access. And this is what has happened in the 21st century with the paradigm shift in scholarly publishing towards Open Archiving. We now have two major providers of scientific information – The Commercial Publishers databases and the Digital Libraries and Archives based on the Open Access Principles and run by Academic libraries and scientific societies.

Apart from the World Wide Web at large the Digital Library must be a controlled zone with a carefully selected subset of information objects which are made searchable and retrievable for the customers/users. The continuing responsibility that comes with a Digital Library is the control of the subset over time and to ensure that it remains stable and accessible. In order to do so you add value to the items moved from outside into the Digital Library. Added value in the form of increased accessibility; in the form of metadata that today is mostly bibliographical but in the future can be used for statistical and rating purposes; added value in the form of maintaining the object for a longer period of time, making sure it is stable, original and can be found at a particular address; added value in the form of peer-review, making sure that the object has been subject to editorial scrutiny; and finally adding value to the object by presenting it in a standardised format that can be viewed, retrieved and preserved and shared according to international standards.

2. Construction and Organisation

2.1. Technical questions

The backbone of any Digital Library is of course computer software, hardware and the reliance on high speed networks. Any Digital Library would rely on components such as local networks with connections to the Internet; databases with user friendly interfaces for searching and administrative purposes which can index full text documents for fast access.

The OPACs (Open Public Access Catalogue) of most libraries are systems that embrace all kinds of content and functions of the organisation in one single system. This is not necessarily true for the Digital Library that most likely is a collection of different systems and resources connected through a network and searched and reached by a web-interface (9).

Since 2002 we have the Open Access Initiative – Protocol for Metadata Harvesting (OAI-PMH) [10] which is widely used by all kinds of institutional repositories archiving scientific material which most of the time fits right in to either two of the definitions of Digital Libraries, and if not they probably are functioning parts of a Digital Library as discussed above. The protocol provides an application-independent interoperability

framework based on metadata harvesting. There are two classes of participants in the OAI-PMH framework: Data Providers [11] that administer systems which support the OAI-PMH as a means of exposing metadata; and Service Providers [12] that use metadata harvested via the OAI-PMH as a basis for building value-added services.
Data providers typically use repository software that are OAI compliant. Today there are several free software tools for creating institutional repositories available [13]. Frequently downloaded, installed and used are programs such as E-prints, D-space, CDSware etc.

2.2. Building the Collection

One of the major issues of Digital Libraries at the moment is the question of creating a digital collection with some sort of scope and critical mass. This is especially true when it comes to offering full-text records. Building the collection usually means four things:

1. Acquiring original digital works created by original authors. This could be articles, books, conference proceedings, pictures etc.
2. Converting paper originals into digital format.
3. Purchasing or using free external material by either incorporating the material, such as electronic books and journals, or creating pointers to external websites.
4. Dealing with rights management issues, such as keeping track of copyright status of the digital material, identifying and authenticating users and their access to the material.

2.3. Document Formats and retrieval

2.3.1. Document Formats

The Digital Library of today uses a wealth of different document formats storing and representing the content. A file format is a software algorithm for encoding the data and any information about the data. It can either be in a proprietary format, which means they are developed and used in software by commercial companies, or there are open file formats that are available for use by anyone.
There are many hundred different file formats in use for different sorts of data. Talking about Digital Libraries and document formats, I will only mention a few important formats which generally can be divided into different categories. One is Plain text formats where we find the most widely used character set encoding ASCII (American Standard Code for Information Interchange) (ISO 641), though it is fast becoming replaced by new and more sophisticated character sets such as UNICODE (ISO 10646-1) designed to cover all the worlds alphabets. Another is proprietary formatted-text formats including MSWord, WordPerfect and other word processing applications. There are also desktop publishing applications like PageMaker, QuarkExpress with their own format coding. Rich Text Format (RTF) is a export/import format developed by Microsoft and used by many word processers.
Page Description formats describe shapes and the layout of a document. They are widely used for presentation of formatted pages in databases which carry journal articles in full

text. The most applied of these formats are the Portable Document Format (PDF) which is a development out of the Post Script (PS) programming language, developed in the Adobe laboratories in the mid 1980s.
Structured information formats do not describe layout, but instead the structure of the document. Standard Generalized Mark-up Language (SGML) and Extensive Mark-up Language (XML) are both meta languages which are used to specify an arbitrary arrangement of mark-up tags to be read by the computer. The arrangement of tags for different elements in the text makes it possible to create self descriptive objects which can be manipulated automatically by a computer system. The markup "grammar" is specified in a so called Document Type Definition (DTD). Different DTDs adapt to different sorts of documents. Encoded Archival Description (EAD) is much used for archival material; Text Encoding Initiative (TEI) is a DTD for humanities texts mainly. One of the most common used DTDs is Hyper Text Markup Language (HTML), which is a reduced tag set version of SGML.
For still images that store information about individual pixels or dots in the picture there are bitmapped formats such as Graphics Interchange Format (GIF). It is limited to 256 colours but is very popular because it displays well on computer screens and creates relatively small files. Portable Network Graphics format (PNG) is intended to replace GIF but has not yet been a public success.
Joint Photographics Expert Group (JPEG) is on the other hand the most used compression format for colour photograph images on the web with over 16 million colour hues available. The Tagged-Image File Format (TIFF) is used as a archival or intermediary format. It stores large amounts of information and creates fields in the mega and gigabyte area. It supports all kinds of compression. For still images that use mathematical algorithms to store information about lines and curves, the standard vector format is Computer Graphics Metafile (CGM).
For Audio and moving image formats the formats of choice lately is proprietary formats such as Audio Interchange File Format (AIFF) which is a Apple audio format. In the other corner we have an originally Microsoft Windows´ audio file format that now also works on other platforms (RIFF WAVE). RealAudio is another popular proprietary audio format which can "Stream" the information. That means the audio file starts to play as soon as the first bits are received by the users computer. Lately the audio format of choice has become MP3. It stands for MPEG Audio Layer 3. It is a compression algorithm for digital audio developed by the Motion Picture Experts Group. MP3s are digital audio files that have been compressed while still maintaining good sound quality. This type of compression enables near stereo quality audio brought down to an extremely small size which made it so popular in the file sharing community.
Multimedia collections of course also include moving image file formats. Many of the most used are proprietary like Apples QuickTime Movies, Windows Audio-Video Interleaved (AVI), Progressive Networks RealVideo and different versions of the compression standard for video – MPEG. In the multimedia classification we also find proprietory interactive formats like Macromedia´s ShockWave used for animated interfaces, advertisement and games production. Also found in multimedia documents is Sun Microsystems Java programming language or Microsoft´s Active X components [14].

2.3.2. Retrieval

Digital Libraries come in many forms. They can contain simple metadata or catalogues of bibliographical information. They can contain the full text of the document. They can contain images, audio or multimedia material. All this information may be available in different formats, created with different software. Most probably the resources reside on different servers using no unified thesauri or heterogeneous indexing schemes. All of this makes information retrieval a very complex task. Every information system is unique when it comes to retrieval methods, and it is more or less necessary to have a fair idea of the characteristic features of each system to be able to perform a relevant search. This becomes even more complex when some Digital Libraries allow users to conduct searches across a range of distributed services [15].

Those who work with human-computer interaction and usability testing have found that users-performance on computer-based systems can be greatly affected by the users´ previous experience and knowledge. This can generally be defined in one of two ways: the knowledge of the search topic or domain, and the knowledge of the system used.

Although it can easily be expected that the domain expert would outperform the novice with little knowledge of the domain, researchers point out that domain specific knowledge begins to predict performance only after users have acquired some experience with the system used. In a study on factors affecting search performance on an on-line database system it was found that system knowledge and computer experience critically influenced search performance.

In another study comparing the effects of users´ search experience and subject expertise on the use of online database systems, it was concluded that the users search experience affected their use of search strategies and played a more important role than did their subject knowledge [16].

So it seems that several studies confirm that both search experience and good search strategy are vital to get a good search result and that these variables mediate subject knowledge. Knowledge affects search strategy and search strategy in turn affects the researcher´s ability to locate relevant information. These findings highlight not only the importance of task-relevant knowledge but also the significance of the role that search strategy plays in the decision-making process.

This knowledge has important implications for both designing more-effective information search aids and training the users of these aids. As the size of the Digital Library and complexity of issues increase, the nature of the search strategy is likely to become increasingly important in explaining the ability to locate relevant information. With very large databases, users are likely to be unable to utilize their knowledge to resolve issues unless they are able to limit the search to areas of the database where the information is located [17].

Web users, on the other hand, show very different patterns of searching from those found in traditional information retrieval systems such as online databases. For example, most users did not have many queries per search session, and each query tended to be short. Boolean operators were seldom used. Many users submitted only one query and did not follow with successive queries. So to support the typical web-searcher when he or she is entering a website that is using traditional IR principals one could recommend to improve the user

interface by including broader resources such as increased browsing and viewing mechanisms and more online help-functions [18].

2.4. Preservation

When it comes to digital preservation and curation we are still living in some sort of digital stone-age. Many a task force on how to archive digital information for long-term presevation have come and gone [19, 20]. Usually their strategy recommendations are founded on infrastructures of common standards, methods and tools. The big problem with preservation research is that it also must deliver practical results. This is often obstructed by reality where organisations use heterogeneous processes for archiving their material with structures that are incompatible for exchange and re-use of resources. But the need for standardized preservation technologies, practices and methods are very much in the public interest. The possibility to retrieve old as well as new cultural heritage material, research material, governmental documents etc. must be available to fortify democracy and to promote economic growth.

One interesting effort to deal with this problem within Digital Libraries was recently made by a working group of the Network for Digital Libraries [21]. They focus on three research challenges that must be addressed if progress of sustainable Digital Libraries is to be made. The areas are:

- Preservation strategies – Emerging research domains that can identify new problems which are the result from constantly evolving technology. For example, we know a lot about media but lack archival media which is sustainable without intervention for more than 100 years.
- Re-engineering preservation processes – Today's preservation processes are slow, costly and often manual. Generally they follow the same roads as preservation practices for physical material. But if preservation processes shall become efficient a complete re-engineering must take place. Automatisation must be a key feature to speed up deliveries and slow down costs. Preservation functionality must be built into the systems used to create and manage the data. This is the only way to guarantee longevity of digital entities.
- Preservation systems and technology – To support this process transformation tools and technologies must be there and answer questions like: How can complex and dynamic entities be authenticated and their integrity verified? How can automatic mechanisms for the creation and authoring of metadata be accomplished? How can we solve issues of multilingualism for searching?

Preservation issues like these insist on answers and the answers must be delivered pretty soon or we will put a great deal of the worlds cultural and science heritage at risk.

2.5. Metadata and Naming

The reason why distances today seem to grow shorter and shorter could be spelled Interoperability. It is a key factor also for the Digital Library. The absolute trend of the 21st century is to adopt standards and develop and implement open systems that support

interoperability. There are of course a multitude of aspects of interoperability but we will here only stop to take a short look on two of these aspects: metadata and naming.
When we search the web we will find a lot of junk but the relevance in the OPAC-search is high. The reason for this fact is that the OPAC-documents are indexed based on a number of document-fields. These fields can be combined or be searched one by one, contrary to the web documents in the Google-search which contain a limited number of tags that can be indexed as fields. To incorporate value added fields into HTML-documents is of course unreasonable both from an economical and human perspective – but it would make life easier...
There is nothing much that can be indexed as a field in an html-document – Title, URL, Link, heading...? Compare that with a MARC-index!
Different persons have thought about this and arrived at one conclusion but offering different solutions. The conclusion is called metadata for web documents. This is simple information about the document which could be added by the creator himself in order to make the Web a better place for those who seek and those who want to be searched out.
For this reason a number of different schemas for web document cataloguing have been created by librarians, archivists and researchers during the last 15 years. One of the more popular schemas is called Dublin Core.
Dublin Core is a simple content description model for electronic resources. It got its name from the small town Dublin in Ohio, USA, where in 1995 representatives from museums, libraries, governments and commercial companies for the first time agreed on creating a core of metadata elements to improve search ability for electronic resources. The elements, it was agreed, should be so simple to use that the author of the documents would be able to catalogue his own resources. The metadata elements should contain a core of information used across most areas of science. The information in these fields must be enough for retrieval and identification of the resource. By using a common set of elements the Dublin Core group wanted to simplify the possibilities for semantic interoperability between different disciplines. Dublin Core has a broad international base for its work since about twenty countries in North America, Europe, Australia and Asia are involved in the developing work. The Dublin Core model is an economic alternative to more detailed description models like MARC and other rules of cataloguing derived from the library world, but it is still sufficient and flexible enough to transpose structure and semantics from richer standards [22, 23].
The different need for metadata on the WWW calls for an infrastructure admitting coexistence of complementary and independent metadata schemas. Therefore the World Wide Web Consortium (WC3) has started to create a metadata-architecture called The Resource Description Framework (RDF) to support interoperability of metadata describing any item that can have a Uniform Resource Identifier (URI). In RDF Dublin Core, for example, is a standard vocabulary in the framework using XML as the encoding syntax [24].
Naming is a problem in the Digital Library. It is vital to have names that uniquely identify digital objects. These names must be part of any documents metadata. Names are important in order to give correct citations; make relevant information retrieval possible; to make links between objects and to manage copyright.

Therefore naming must be permanent and can not be tied to a specific location. The name and the location must be separated. This is just the opposite to the current method used for identifying objects on the Internet today where we find information about the method by which a document is accessed. The machine name and the document path and the file which harbour the specific document is included in the same string in the Unique Resource Locator (URL). In this system when the file is moved the document is lost. That is why so much work has been done to find a scheme of unique identifiers that have persistence beyond the life of the server or the organisation. There are several of these schemes that have found solutions for giving documents persistent names that are valid whenever documents are moved from one location to another, which is often the case due to administrative reasons or migration from one medium to another.
Three major schemes have developed lately and they all are built around the idea to separate a document name from its location by mapping information of a unique never-changing name to one or more URLs. They all are based on the assumption that an institution, like a national library, must take the responsibility for managing such a system.
The most popular systems are Uniform Resource Name (URN) developed by the Internet Engineering Task Force (IETF); Digital Object Identifier (DOI) initiated by the Asssociation of American Publishers and the American Corporation for National Research Initiatives; and Persistent Uniform Resource Locator system (PURL) developed by Online Computer Library Center in Ohio [25].

3. Digital Library types and content

The Digital Library community is clearly increasing in number and volume as more and more people get connected to high speed internet connections, more people get involved in distance learning, more people get used to online communication, governments, institutions and commercial companies realize the potential in digital deliveries. Developments like these have prepared the ground for a large number of different types of Digital Libraries throughout the world. It is difficult to classify a phenomenon with a definition still under debate but for the purpose of this short introduction I would like to group Digital Libraries into 5 types and give a few examples of each:

- Digital Libraries at scientific societies or organisations
- Digital Libraries at Commercial publishers
- Digital Libraries at National Libraries
- Digital Libraries at Universities
- Digital Libraries at Museums and other cultural heritage organisations

3.1. Digital Libraries at scientific societies or organisations

The Institute of Electrical and Electronics Engineers IEEE provides access to almost one third of the world literature in the area of electrical engineering and computer science. Their Digital Library called IEEE Xplore provides full-text access to IEEE transactions, journals, magazines and conference proceedings published since 1988 and all current IEEE Standards. IEEE Xplore covers technical areas ranging from computer engineering,

biomedical technology and telecommunications, to electric power, aerospace and consumer electronics. For full access you have to be a member (http://ieeexplore.ieee.org/)

Association for Computing Machinery (ACM) members and registered users can use the Digital Library containing bibliographic information, abstracts, reviews, and the fulltext for articles published in ACM periodicals (journals, magazines and transactions) and ACM proceedings (http://portal.acm.org/).

Close to one million documents of interest to people working in particle physics and related areas can be found at the CERN document server site. Originally named Conseil Européen pour la Recherche Nucléaire, now renamed European Organization for Nuclear Research, the Digital Library at CERN covers preprints, articles, books, journals, photographs, and much more available at no cost for everyone (http://cdsweb.cern.ch/).

BioMed Central is an independent publishing house committed to provide immediate free access to peer-reviewed biomedical research. BioMed Central publishes more than 50 on-line journals covering the whole of biology and medicine. The service includes support for journal editors in developing countries (http://www.biomedcentral.com/).

3.2. Digital Libraries at Commercial publishers

The content of these Digital Libraries are mostly the same as of Digital Libraries of scientific societies and organisations – bibliographical or full text copies of journal articles, conference proceedings etc. drawn from a single or distributed databases. Major examples of this type are:

Springer Link, with a collection of journals and book series that account for over 1 million documents in different kind of subject areas from Springer Science Publishers (http://www.springerlink.com/).

ScienceDirect which is the giant publisher Elseviers collection of science, technology and medicine full text and bibliographic information of the same kind as you find in Springer Link (http://www.sciencedirect.com/).

ISI-Web of Knowledge accesses multidisciplinary databases of bibliographic information gathered from thousands of scholarly journals. The databases are indexed so you can search for specific articles by subject, author, journal, and/or author address. Because the information stored about each article includes the article's cited reference list you can also search the databases for articles that cite a known author or work (http://portal.isiknowledge.com/)

3.3. Digital Libraries at National Libraries

National libraries are the collective memories of nations. They always house valuable collections of both scientific and cultural nature where citizens go for research and investigation of historical heritage and events. These days many national libraries take advantage of modern technology to help them serve their customers need. By digitizing text, sound, film-collections and making them into Digital Libraries within the national library, cultural heritage treasures are made available for people far outside the national borders. Fine examples of this are to found at:

British Library which offers a number of digital information services based on British Library collections. One example is the "Treasures in Full" site that brings high-quality editions of the works of Shakespeare, Chaucer and the Gutenberg bible among other things to the users desktop (http://www.bl.uk/treasures/treasuresinfull.html). Another is "Images online" which gives access to thousands of pictures from the library collections (http://www.imagesonline.bl.uk/britishlibrary/).

Library of Congress offers two great Digital Library resources: "American Memory" which is an umbrella term for a collection of digital resources on topics such as African American History, Immigration, Native American History, Performing Arts etc. (http://memory.loc.gov/ammem/) and a resource called "Thomas", named after Thomas Jefferson, made up of several databases, where federal legislative information is freely available to the Internet public (http://thomas.loc.gov/).

The National library of Portugal can stand as an example of a smaller national library with limited resources that is working towards a vision of a National Digital Library consisting of a coherent group of services and resources with technical solutions based on open and scalable technology. Several digital collections and exhibitions are available. A good example is the digital collection devoted to the renaissance portugeese mathematician and astronomer Pedro Nunes (http://purl.pt/40/1/) [26] (Note the persistant URL!).

3.4. Digital Libraries at Universities

In the second half of the 1990s several university libraries started building digital collections making them public available.

The Electronic Text Center at the University of Virginia is one famous Digital Library offering thousands of SGML-encoded electronic texts and many special collections devoted to famous authors or American historical events (http://etext.lib.virginia.edu/).

Project Gutenberg also emanated from a university. In this case the University of Illinois. The objective of the project was to provide free access to digital version of world literature. The texts are stored in plain ASCII format making them easy to read and search with any sort of computer equipment. Today Project Gutenberg is a volunteer effort continued outside the university (http://www.gutenberg.org/).

At the Oxford Digital Library a number of disparate collections from the university are available. You can go from images of medieval manuscripts to a database of Athenian Pottery to a collection of motoring and transport images. In the future the Oxford Digital Library aims to offer a Digital Library architecture which will allow centralized access to these digital resources. The use of established standards for descriptive metadata (i.e. EAD,

TEI) is a precondition for this integration process. Existing Digital Library collections may be transferred step by step into a common architecture with an integrated retrieval mechanism (http://www.odl.ox.ac.uk/).

Established in 1997 as a University of California library, the California Digital Library has become one of the largest Digital Libraries in the world providing access to resources like The Online Archive of California – collections of digital materials (such as manuscripts, photographs, and art) held in the libraries, museums, and archives across California; government data and statistics about California in "Counting California"; The Melvyl catalogue with its 15 million records from the 10 University of California campuses. (http://www.cdlib.org/).

arXiv is an e-print service in the fields of physics, mathematics, non-linear science, computer science, and quantitative biology that started in 1991. It is a fully automated electronic archive and distribution server for research papers The contents of arXiv conform to Cornell University academic standards. arXiv is owned, operated and funded by Cornell University, a private not-for-profit educational institution. arXiv is also partially funded by the National Science Foundation. Users can retrieve papers from the archive either through an world wide web interface, or by sending commands to the system via e-mail (http://arxiv.org/).

3.5. Digital Libraries at Museums and other cultural heritage organisations

Many cultural heritage institutions are building digital collections of their holdings in order to provide easy and affordable access to the cultural heritage resources. In many ways it has been a tough trip since funding and research for Digital Libraries of this type have not been as excessive as in other areas. There is also the problem of common standards in order to describe cultural objects homogenously, using the same kind of metadata standards. Cultural heritage institutions use a multitude of different standards and in many cases no standards at all. This will be one of the big challenges for the next generation of Digital Libraries at museums and cultural heritage sites [27].

The State Hermitage Museum in St. Petersburg, Russia has provided access to many collections of the museum. A variety of techniques have been used for making 3D images and virtual exhibitions come alive in both Russian and English (http://www.hermitagemuseum.org/)

The 24 Hour Museum is the UK's National Virtual Museum, offering a unique mix of dynamic content including daily arts and museum news as well as exhibition reviews. It functions as an access point to Cultural heritage sites of the UK. Venue and listings info is driven by a comprehensive searchable database of more than 3,400 entry points.

This brief expose hopefully has made clear how varied and heterogeneous the flora of Digital Libraries are today. Some work with very simple technology, like Project Gutenberg, but others like virtual worlds of the Hermitage are very sophisticated. Some are designed to provide access to digital resources in specific fields like BioMed Central or arXiv. Others give access to specific document types in a wide area of scientific subjects provided by commercial publishers that charge you for the data, while many other Open Access archives are available for free [28],[29].

4. Future Developments

As we have seen Digital Libraries are thriving and expanding their services worldwide not caring too much if they fit into one definition of some sort or the other. Many have been sponsored and funded by government bodies such as the eLib Programme in the UK [30] and the Digital Library Initiatives phase 1 and 2 in the United States [31].

What is being done at university libraries, at commercial publishers, scientific societies, museums or any other organisation producing and managing information for a community of users, is much more than converting analog data into digital form. New material is created and served in new forms. Take for example how resources from different collections are put together into new entities of information by making great use of pictures and sound and all kinds of different search facilities, which is just impossible, for economical or technological reasons, to bring about in the analogue world. These resources go side by side with full-text journals, books, references available at your desktop in seconds. Just click on the link! There we go again – the ghost of Vannevar Bush´s Memex. But we are still far away from a perfectly working Memex kind of gadget. There are several stumbling blocks that still irritate the everyday user and the visionary when using the Digital Libraries of today.

We have, for example, the technical issues which include the problem with standards and protocols. To bring the distributed variety of digital resources and services together in a way that allow for integration and unified search, retrieval and presentation is a great challenge for the future. So is the problem of transferring personalised service and support from standard library and information services to the Digital Library. A user interface can hardly replace person to person service but better user interfaces must be developed and researched in order to help users. The future Digital Library will go beyond helping the user with searching and browsing only. Users must be able to expect support for taking correct actions and getting help for problem solving where the Digital Library system confirm or deny existing hypotheses. In an interesting paper L. Feng et al.[32] distinguishes between traditional searching and browsing which is called “tactical level cognition” and the problem solving act which is called “strategic level cognition”. In the future, the authors argue, Digital Libraries must become not only a simple storage place but a place where knowledge is acquired, shared and multiplied. To facilitate the browsing function Digital Libraries must integrate diverse repositories of coherent collections and include navigation, searching and browsing facilities in a network of inter-related concepts and repositories. That takes care of the “tactical level cognition”. The “strategic level cognition” support must provide justifications and evidences by adding value and advocating a closer interaction between users and the content. To do this it is necessary to use some sort of vocabulary. Parallel to classic keyword-based indexes, knowledge-based index must be constructed. The authors outline a framework for a machine centred and extracted knowledge discovery across multiple repositories. This is being done in six steps by setting up knowledge discovery targets; identifying relevant resources; filter out interesting concepts; correlating concepts; extracting knowledge and justifications from correlated concepts and the evaluation of the same.

Well, this is exactly in the Vannevar Bush spirit! It is one possible road ahead. But before we are there the solution to everyday problems like long term preservation of digital

objects; copyright of digital material; good solutions for microcharging and pay per view; how to bridge the digital divide and include and promote Digital Libraries of developing countries are maybe more imminent and real. They must and will be solved!

More and more digital material is added every day to the web. We are just learning how to deal with it in the best way. We are only in the beginning of a long and winding road. One of several future stops on that road is called the semantic web. It is a vision trying to remedy the Babel problem of today´s web by machine processable language ontologies. Ontologies provide a shared understanding of a topic of interest among humans and computers. The mere mass of information that is added every second to the Web, calls for machine processability. How can the future semantic web help Digital Libraries? The simple answer is that if we had common schemes in the form of ontologies helping us naming and cataloguing digital objects this would enable interoperability. The user would think he is navigating one single Digital Library system but in reality he would be using a multitude of distributed systems.

By creating standard machine processable ontologies, ontology editors, annotation tools and inference engines that deduce new knowledge from already specified knowledge (as outlined above by Feng et al) it will be easier in the future to add semantic markup and metadata to documents making them not only richer in content but also much easier to get hold of [33].

Content management technologies will be the big thing of the future. The increasing amount of digital content will see to that. Semantic web technologies will probably add important features to Digital Libraries like semantic interoperability, better browsing, searching and filtering capabilities and delegating routine tasks of cataloguing, metadata annotation etc to automated agents [34]. This is not going to be an easy task. The process of creating and administrating quality metadata records based on shared and ever evolving ontologies is a heavy one but the stone has already started rolling and sooner or later it will roll past a library near you.

Maybe it is metadata that describes and finds content and relationships between content that will be the infrastructure of the future Digital Library. Maybe that´s the sort of stuff that will make Vannevar Bushs vision, of a world of information available by your fingertips at your desk in your own office or at home, come true. Who knows? One thing is for certain, though. The Digital Library, in whatever future shape it may take, is here to stay.

References

[1] Vickery, Brian C. Scientific Communication in History. Scarecrow Press Inc. 2000. Page 81-111.

[2] Bush, Vannevar. As We May Think. Atlantic Monthly, July 1945. pp.101-108.

[3] Chowdhury, G.G. and Chowdhury, S. Digital Library research: major issues and trends. Journal of Documentation, 1999. Vol. 55(4), pp.409-448.

[4] Borgman, Christine L. What are Digital Libraries? Competing Visions. Information Processing & management, 1999. Vol. 35(3) pp. 227-243

[5] Oxford English Dictionary. http://dictionary.oed.com/

[6] Witten, I.H. and Bainbridge, D. How to build a Digital Library. Morgan Kaufmann. 2003

[7] Waters, D.J. What are Digital Libraries? 1998. CLIR (Council on Library and Information Resources) Issues, No.4. http://www.clir.org/pubs/issues04.html.

[8] Atkinson, Ross. Library Functions, Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the Control Zone. The Library Quarterly, 1996, Vol. 66(3), pp. 239-265

[9] Cleveland, Gary. Digital Libraries: Definitions, Issues and Challenges. UDT Occasional Paper #8. March, 1998

[10] OAI Protocol for Metadata Harvesting: URL http://www.openarchives.org/OAI/openarchivesprotocol.html

[11] OAI Registered Data Providers. http://www.openarchives.org/Register/BrowseSites

[12] OAI Registered Service Providers. http://www.openarchives.org/service/listproviders.html

[13] Open Archives Tools. http://www.openarchives.org/tools/tools.html

[14] FOLDOC – Free on-line Dictionary of Computing. http://wombat.doc.ic.ac.uk/foldoc/index.html

[15] Chowdhury, C.G., Chowdhury, S. Introduction to Digital Libraries. Facet Publishing, 2003

[16] Palmquist RA, Kyung-Sun K. Cognitive Style and On-Line Database Search Experience as Predicitors of Web Search Performance. Journal of the American Society for Information Science, 2000. Vol 51(6) pp. 558-566

[17] Barrick JA, Spilker BC. The relations between knowledge, search strategy, and performance in unaided and aided information search. Organizational behaviour and human decision processes, 2003. Vol. 90(1), pp.1-18

[18] Wolfram D, Xie H. Traditional IR for web users: a context for general audience digital libraries.Information processing & management, 2002. Vol. 38(5), pp. 627-48

[19] Waters, G; Garrett, J. Preserving Digital Information: Report of the Task Force on Archiving of Digital Information. The Commission on Preservation and Access and the Research Libraries Group http://www.rlg.org/ArchTF/

[20] A Strategic Policy Framework for Creating and Preserving Digital Collections [Copyright Internet Scout Project, 1994-1999. http://scout.cs.wisc.edu/] http://ahds.ac.uk/strategic.htm

[21] Ross, Seamus; Hedstrom Margaret. Preservation research and sustainable digital libraries. Int J Digit Libr, 2005. Vol. 5(4), pp. 317-324

[22] Introduction to Metadata: Pathways to Digital Information. Edited by Murtha Baca. 1998 The J. Paul Getty Trust. ISBN 0-89236-533-1. Available at the web: http://www.getty.edu/research/institute/standards/intrometadata/

[23] Milstead, Jessica and Feldman, Susan, Metadata: Cataloging by Any Other Name. http://www.onlinemag.net/OL1999/milstead1.html

[24] Resource Description Framework. http://www.w3.org/RDF/

[25] Cleveland, Gary. Digital Libraries: Definitions, Issues and Challenges. UDT Occasional Paper #8. March, 1998.

[26] Borbinha, José Luis. An approach to creating a national Digital Library. Int J Digit Libr, 2004. Vol. 4(1), pp. 19-22

[27] The DigiCULT Report. Technological landscapes for tomorrow´s cultural economy Unlocking the value of cultural heritage. 2002. European Commission Directorate-General for the Information Society. ISBN 92-828-5189-3.

[28] Lesk, Michael. Understanding Digital Libraries. Second Ed. Morgan Kaufmann Publisher, 2005

[29] Tedd, Lucy A; Large, Andrew. Digital Libraries. Principles and Pracitce in a Global Envrionment. K.G. Saur, 2005

[30] eLib: The Electronic Libraries Programme http://www.ukoln.ac.uk/services/elib/

[31] Digital Libraries Initiative phase 2. http://www.dli2.nsf.gov/

[32] Feng, Ling; Jeusfel, Manfred A; Hoppenbrouwers, Jeroen. Beyond Information Searching and Browsing Acquiring Knowledge from Digital Libraries. Information Processing and Management, 2005. Vol. 41(1), pp. 97-120

[33] Sure, York and Studer, Rudi. Semantic Web Technologies for Digital Libraries. Library Management, 2005. Vol. 26(4/5), pp. 190-195

[34] Lytras, Miltiadis; Sicilia, Miguel-Angel; Davies, John; Kashyap, Vipul. Digital Libraries in the Knowledge Era : Knowledge Management and Semantic Web Technologies. Library Management, 2005. Vol. 26(4/5), pp. 170-175