managing
information

aslib
MANAGING INFORMATION

for everyone who uses information

Research Reviews:

Issues in digital libraries
- E-books about Central Asian Countries
- Usability and Usefulness
- Automatic document analysis
- The next 250 years: Ensuring long term learning with the British Library's Digital Library

Cyberbullying and other online risks

Internet future: Issues to watch

NEW Regular Columns!
- Copyright Surgery
- Technologically Speaking
- Career Coach

Pulling Rank: Bibliometrics for assessing impact

BMA Library 21st century make-over

Internet Librarian International 2011 Preview

managing
information

Contacting ASLIB

Further information about ASLIB, its activities and services, can be
obtained from:

ASLIB
Howard House
Wagon Lane
Bingley
BD16 1WA
United Kingdom

www.aslib.com

https://twitter.com/ASLIB_info

Tel: +44 (0) 1274 777700
Fax: +44 (0) 1274 785201

aslib
MANAGING INFORMATION

ASLIB is a membership association for people who manage information and knowledge
in organizations, who are not necessarily librarians.

It was originally established as the Association for Special Libraries and Information
Bureaux in London, in September 1924. Our original ethos, which still applies today, was
"*to serve those engaged in the collection, treatment and dissemination of information in
many departments of human activity.*"

In April 2010 ASLIB was acquired by MCB Group, the holding company for Emerald
Group Publishing Limited. The acquisition means that ASLIB's services to members will
continue and grow in line with member and community feedback.

ASLIB provides training, advice and networking for members, with a key focus on data
protection, intellectual property and information governance issues.

ASLIB members receive the *Managing Information* magazine 10 times a year, plus
a choice of other professional and scholarly publications dealing with information
management and wider business and management issues.

Contact: members@aslib.com

The ASLIB Team:

ASLIB Publisher: Diane Heath - **dheath@aslib.com**

ASLIB Relationship Manager: Holly Shukla - **hshukla@aslib.com**

ASLIB Director: Rebecca Marsh - **rmarsh@emeraldinsight.com**

Managing Information Editor-in-chief: Graham Coult - **gcoult@aslib.com**

Welcome to Managing Information

Volume **18** Issue **7** 2011

Managing Information magazine - YOUR eye on the information world

Managing Information is a subscription magazine available both electronically and in print format, for everyone who uses information.

The magazine's core topics include information management, knowledge management, taxonomy, intellectual property (copyright, patents, trademarks, data protection, freedom of information), IT security, disaster recovery (information and data aspects), the internet, Web 2.0 and social media, libraries, information centres and museums.

Managing Information reflects the needs of information professionals across all sectors - corporate and commercial, public, academic/educational, and voluntary. The purpose of the magazine is to help readers enhance their own performance, and that of their organizations, providing news, comment and analysis, feature articles on best practice, and reviews.

Online access:

Contact **support@aslib.com** for login details.

www.managinginformation.com

Contents

Research Review:

Issues in digital libraries

E-books about Central Asian Countries – producers and repositories

Introduction

E-books are becoming ever more common. Open access is also a very popular trend. Open access to information sources means their availability, free-of-charge to the user, on the public Internet allowing users to read, download, distribute, print, or link to the full text of the materials concerned. Users may also crawl them in order to create indexes and make use of them also for other purposes. This should be achievable without financial, legal or Technical barriers.

Open access is seen as removing barriers arising from price and the need to seek permission in order to make copies.

A piece of research entitled "Open access e-book collection on Central Asia in selected digital archives" looked at the size and quality of e-book collections available in Central Asia. Is there a satisfactory collection of open access e-books on Central Asia? Do collections include made digital and born digital materials? Where were most books published? What topics were covered? This piece of research aimed to find out.

Online publishing – the early years

Online publishing as an idea is not as new as we might think. In 1945, there was a vision of a device called a "Memex" which was able to store, retrieve and display books, records and documents. The key to this vision was that the Memex should be mechanized so that it could operate very rapidly. It led eventually (at least through the aspirations which led to it being thought up) to the development of the World Wide Web and hypertext.

Another important development, Project Gutenburg, is also not as recent as we might think. It began forty years ago in 1971. A student at the University of Illinois (Michael Hart) was given (almost unlimited) access to a computer which was one of the 15 nodes which would later make up the beginnings of the Internet. At that time he digitized the US declaration of Independence. He wanted to contribute something of value and set the objective of making the 10,000 most consulted books available to the public at little or no charge by the end of the 20th century.

The HathiTrust's activities also constitute another milestone in the development of digitized open access. This project brings together major institutions and libraries to preserve the cultural record for use in the future, it has digitized more than 8 million volumes so far. A further development is the Internet Archive (www.archive.org) which started out with the objective of preserving web pages, but subsequently expanded to include books, films and music. The National Academies Press, another source, was inaugurated in 1996 offering free-to-access digital books while charging for the print versions of the same titles.

Gradually therefore, e-books and digitized collections have grown in importance, just as the popularity of Open Access has grown.

Big is bountiful

The Internet as a resource is now enormous. This piece of research narrowed its focus down to material relating to the social sciences, and also to the largest continent on earth, central Asia.

The research found that there are 170 open access e-books available in databases, archives and libraries relating to Central Asia. The highest number of these 170 is available on Google Books (31.8%). The next highest number of the e-books are available courtesy of the US Strategic Research Institute (24.71%). Next comes Human Rights Watch – 19.4% of the e-books are available through its databases.

The extent to which e-books are downloadable from the repositories under study was also looked at. 100% of the open access books available at the Strategic Research Institute in the USA were found to be downloadable. The same was true at the Asian Development Bank, Central Asian Caucasus Institute (in the US), and the Council of Foreign Relations. However, at Human Rights
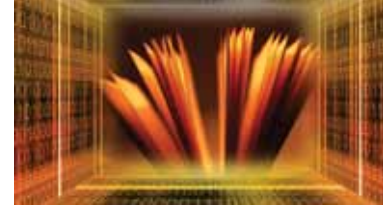
Watch only 75% of their e-books collection was downloadable. This fell to 62.5% at the California Digital Library, 57% at the World Health Organization – and only 5.66% of the e-books on Google books are downloadable.

Subject matter and focus

In terms of geopolitical focus, 88 e-books were concerned with 5 republics of Central Asia – Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan and Uzbekistan – while 30 are about Afghanistan, 19 focus on Ladakh and Kashmir, 10 on Xinjiang and Tibet, 8 on Iran and the remainder cover other parts of Central Asia.

In terms of subject matter, 32.35% are concerned with politics of Central Asia, while 15.29% cover its sociology, 9.41% deal with its economics, and 7.65% its history. The research also found that 24.71% of the e-books were multi-disciplinary, and 7.65% deal with other aspects such as religion and education.

Country of publication

Most of the e-books were published in the USA (73.53%). The UK accounted for 8.24%, Denmark 4.12%, the Philippines also 4.12% and France published 2.94%. Other countries include Japan, Argentina, Turkey and Sweden.

Good collection but needs work on awareness

The researchers took the view based on their findings that there is a good collection of open-access e-books available via the Internet about Social Sciences and also Central Asia. They have identified a need though for researchers to be made aware of open access e-books materials in digital repositories, institutional repositories, digital archives, digital libraries databases etc. It is recommended that awareness be raised through workshops, orientation courses and literacy programmes. Library and information professionals are seen as having a key role in educating users on how to find e-books, read them and make use of them. In order to be able to do this, librarians specializing in digital materials must get out onto the Internet and research, identify the best resources in a given topic, and put their users in touch with those resources they find which are relevant to the user's needs.

Comment

This review is based on "Open access e-book collection on Central Asia in selected digital archives" by Fayaz Ahmad Loan. The Internet has become a very rich resource for materials, not least in the Open Access sphere. However, precisely because of the sheer weight of material available, valuable information can often be lost, or at least overlooked. It is interesting therefore to read a study of what is available with a particular geopolitical focus. Although it is to be expected that the US would be a significant producer, the extent of its dominance is striking, as is the comparatively low level of the next biggest supplier, the UK. As in many areas of information provision today, training and awareness are key issues and pressing needs.

Reference

Loan, F. A. (2011), "Open access e-book collection on Central Asia in selected digital archives", *Collection Building*, Vol. 30 No. 3, pp.126 - 130. ISSN: 0160-4953.

Usability and Usefulness:
combined measuring to assess value

Introduction

System usability is an important consideration when developing a digital library, but what about system usefulness. Is there any point in having a usable system if it is not also useful? Is measuring the two together, usability and usefulness, likely to lead to greater user satisfaction and usage? If that is the case, what should be measured to assess usability and usefulness, and how should it be measured? A piece of research carried out by Steven Buchanan and Adeola Salako of the Department of Computer and Information Science at the University of Strathclyde aimed to find out.

What are usability and usefulness?

Usability relates to the interaction between the human and the computer. Usefulness relates to whether the system supports user activity. The differences between the two are similar to the difference between form and function.

So what are the attributes of usability? There is an ISO standard, ISO 9126-1[1], which sets out specifications for understandability, learnability, operability and attractiveness. Another ISO standard, ISO 9241-11[2] sets out specifications for effectiveness, efficiency and satisfaction. Various authors have been identified as a result of the research who also put forward various aspects of usability and usefulness. While the attributes set out by standards and authors may differ, the researchers have found that the interpretation is closely associated.

Out of the study of standards and appropriate works by authors in the
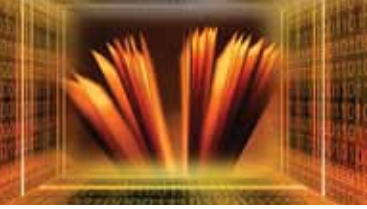
research community, the following attributes were selected to be the focus of the research:

- Effectiveness;
- Efficiency;
- Aesthetic appearance;
- Terminology;
- Navigation;
- Learnability.

These attributes were then subject to further refinement and more detailed definition. Once the key usability and usefulness attributes had been identified, an integrated measurement framework was developed and then an appropriate methodology was devised. A pilot study then focussed on an interactive search system which had been set up a health service as part of its e-library service.

Pilot study - clinicians

Clinicians participated in a questionnaire, and out of those completed, approximately half were by those working in nursing, midwifery and hospital medicine roles. The remainder were evenly distributed across general practice and the allied health professions. Of those responding, 3.4% were found to spend no time online, while 82.6% spent between 5 and 9 hours each week online. The remaining 13.8% spent more than 10 hours per week online.

Overall, the satisfaction levels were found to be positive, although there were some issues identified relating to efficiency, terminology, navigation and relevance. With regard to efficiency, some of those who responded said that:

- response time was too slow,
- too much time and effort was required to enter passwords repeatedly,
- filtering terminology was too time-consuming,
- some terminology was difficult to understand,
- some of the labelling was obscure,
- the system was too complex.

Some respondents also said that they experienced irrelevant results. Despite the above negative comments, 86.2% said they would use the system again.

Five clear recommendations emerged out of the questionnaire on how to improve the system. These were:

- increased use of colour to guide interaction,
- provision of an online guide (especially for constructing search queries),
- single log-on,
- retrieved documents presented in order of relevance,
- increased awareness of the system among staff.

Pilot study – observation of searches

A process of observation of the clinicians with a narrower group participating was also undertaken. Following the observation period it was found that participants were very frustrated with repeated log-on for various sites, they experienced some difficulty in constructing search queries, some used more unstructured natural language for queries and they failed to obtain results and when using the 'browser back' option to return to the portal homepage became disorientated.

After the period of observation the participants said that the navigation was not straightforward and that the system seemed to be slow, although they thought it possible that this might be network rather than system related.

It was thought that the irrelevance of retrieved documents identified by one participant might have been due to over-broad terms might have been used to index the documents. Again, the participants highlighted the need for single log-on, and also said that summaries of documents retrieved, sample queries and increased error tolerance (e.g. misspelled terms). There was much comparison with Google. It was found that the purpose of the system was not apparent, with some questioning its function in relation to the existing e-library service.

The aesthetics of the system received favourable comment but were not seen to be key attributes.

The in-house system was then compared to the commercial alternative and scored better in these categories:

- aesthetics,
- currency.

The in-house system was perceived as similar compared to the commercial alternative in these categories:

- efficiency,
- relevance,
- reliability.

It was perceived as worse under these headings:

- terminology,
- navigation,
- learnability.

Of the participants, 75% expressed a preference for the commercial alternative system.

What and how to measure – the findings

The study set out to deal with a crucial challenge to adopting a unified approach to testing system usability and usefulness – to identify what to measure and how to measure it.

In terms of what to measure, the following attributes have been identified as key attributes of system usability by the research:

- effectiveness,
- efficiency,
- aesthetic appearance,
- terminology,
- navigation,
- learnability.

The following meanwhile have been identified as key attributes of system usefulness:
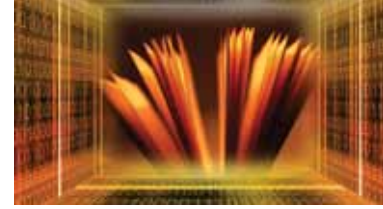
- relevance,

- reliability,

- currency.

In terms of how to measure, the research has found that usability and usefulness can be combined, and that questionnaire and observation are valid approaches, It is however recommended that further research be undertaken in a variety of conditions using more combinations of methods and larger samples. The researchers also found that usability and usefulness are not merely related

properties, they are also dependent properties of system satisfaction.

Comment

This review is based on "Evaluating the usability *and* usefulness of a digital library" by Steven Buchanan and Adeola Salako. Measurement and metrics are important aspects of modern practice, not least in the library and information sphere. Given the level of investment in computer infrastructure and systems required for an effective digital library, it is important that accurate measures of the services provided

are undertaken. Usability and usefulness are key attributes of a digital library which adds significant value justifying the investment made. This study has useful findings to support the effective evaluation of digital library systems, so that they can be refined and improved for the benefit of users, and ultimately the whole organization.

References

Buchanan, S. and Salako, A. (2009), "Evaluating the usability and usefulness of a digital library", *Library Review,* Vol. 58 No. 9, pp.638 - 651. ISSN: 0024-2535.

1. International Organization for Standardization (2001), *ISO/IEC 9126-1: Software Engineering, Product Quality, part 1: Quality Model*, Geneva: International Organization for Standardization.

2. International Organization for Standardization (1998), *ISO/IEC 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability*, Geneva: International Organization for Standardization.

Automatic document analysis:
Ireland's digital library of syllabi

Introduction

There has been a considerable growth in the amount of electronic materials associated with academic courses available via the Internet and intranets – for example syllabus documents and lecture notes. This results in a need for reliable central repositories for such documents so that educators and learners can conveniently share, search and access these materials. In Ireland, a National Repository for such documents has been created and a piece of research undertaken by Arash Joorabchi and Abdulhussain E. Mahdi of the Department of Electronic and Computer Engineering at the University of Limerick, Ireland, looks at how this has been achieved. What were the main elements of this new repository, and how has it all performed?

A syllabus is essentially a summary of a course or a module of study and it provides information on various aspects of the course such as aims, objectives, learning outcomes, schedule, recommended resources and assessment methods. Such documents are valuable not least because they act as the initial contact point between a student and their instructor or tutor. They are also seen as

constituting a form of agreement between the student and the educational institute embracing such items as expectations of required prior learning, topics covered, assessment, qualification, regulations and policies.

The lack of a centralized repository for syllabus documents led to inefficient storage. It meant that syllabi stored were often out-of-date, and the re-use of existing syllabi was often prevented. It was decided that a national structured repository was needed which would hold the syllabi of the majority of higher education courses in the country. A high degree of automation was seen as desirable.

A National Syllabus repository

A prototype repository was developed within the national IDEAS project which is lead by the University of Limerick. The aim is to create an information portal to support the Continuing Professional Development of non-traditional students in Ireland up to and including postgraduate, masters and professional doctorate levels. The portal is intended to provide an expert self-assessment system to assess qualifications and

experience and then offer customized advice on acquiring the necessary knowledge and their career progression. The development of a national syllabus repository in order to supply information on courses and modules supports the work of the IDEAS project.

Because of the nature of the syllabi, with complex documents which are read by humans rather than machines, challenges were posed. There was however a need for automatic syllabus collection for the project to be sustainable, and tools such as web-crawlers and special search engines were considered.

Document classification

Classification of the documents gathered is also important. Classification is at the very heart of library expertise. There are advantages in using a pre-determined classification system such as the Dewey Decimal and Library of Congress systems, but on the other hand the manual classification process is time consuming and can often be tedious and requires expertise. Automated systems were considered, but none was found which was totally adequate to classify the collection of syllabi. An interim compromise had to be found.

> **The portal is intended to provide an expert self-assessment system to assess qualifications and experience and then offer customized advice on acquiring the necessary knowledge and their career progression.**

All these considerations were taken into account when the repository was being developed. A system was devised which consisted of a metadata generator made up of a pre-processing unit, an information extractor, a classifier and a post processing unit. A hot folder application which communicates with an FTP server was also added in to allow individuals and institutions to add syllabi to the system – authorized users simply drag-and-drop their documents onto the system.

The processing system

Once the package of syllabus information has been uploaded to the FTP server, the contents are scanned and processed at regular intervals by the meta-data generator.

It was originally assumed that syllabus documents would be accepted in any format, however it was subsequently found that the vast majority come as PDF or Word documents. To simplify processing, it was decided to convert all documents to a portable format – for the time being this is PDF. The Pre Processing Unit of the system checks the FTP server every second. It transfers any newly uploaded documents and then puts them in a queue to be handled one at-a-time on the repository's main server. Non-PDF files are converted, and then all PDF files in turn are changed to pure text using the Xpdf application preserving as much of the original layout as possible.

The Information Extractor element then takes over. Firstly, it breaks down the document into segments describing specific aspects of the courses. It uses a predetermined thesaurus of terms which are likely to arise in a syllabus to analyze the documents and pick out course modules etc.

Once this process is complete the document is handed over to the Module Syllabus Segmenter. The purpose of this element is to extract topical segments

from the syllabi – again using a thesaurus. When this task is complete the document goes to the Named Entity Extractor for the final stage of information extraction. The segmentation helps to reduce the scope of search for this element of the process, which improves accuracy. The Named Entity Extractor looks for syllabus-related named entities such as module name and module code. Once this process is over, the outputs are passed to the post processing unit for the creation of new syllabus records.

The document then goes to the Classifier. This automatically assigns a classification code to each individual course/module using the ISCED classification system. It has been recognized that a greater degree of accuracy and detail is necessary than the ISCED scheme provides, so the Irish Higher Education Authority is considering the development of a bespoke standard classification scheme.

Once the classification process has been gone through, the document passes to the Post Processing element. This stores the generated metadata for each individual module syllabus along with a copy of the original document as a new syllabus record in the repository's relational database. It uses the information produced by the previous elements to fill in the fields of the syllabus records.

Results of the pilot and lessons learned

A total of 200 syllabus documents from five institutions have been processed during the pilot phase. A number of issues have been identified. The module name and codes in 22 of the documents appeared in large fonts at the beginning of the documents and were not extracted – this led to a 3.7% decrease in the recall of named entities.

Named entity prefixes which did not appear in the thesauri used by elements in stages

of the syllabus processing meant that these elements were not extracted. This meant that 17.6% of potential named entities, and 13.6% of potential topical segments were not recalled.

In some cases where named entities were produced in tabular form, it confused the elements of the system which processed the information, and 5.61% of the potential named entities were not recalled.

Where an unidentified segment follows an identified one, the system assumes the unidentified entity is part of the identified one, and this led to a decrease in accuracy too.

As a result of the project and the experience gained from the pilot, work is now proceeding on ironing out the problems identified above. For example a facility for table detection is to be added in order to help improve accuracy.

Comment

This review is based on "An automated syllabus digital library system for higher education in Ireland" by Arash Joorabchi, and Abdulhussain E. Mahdi. Although this project is concerned with the creation of a national syllabus repository, it is of interest to anyone who is seeking to automate the analysis and processing of particular types of documents. The section on the classification tools looks at the topic in some depth, along with the processes which have been devised, and is well worth reading. The issues which have been identified, and the solutions which have been devised show the benefits of a pilot project.

Reference

Joorabchi, A. and Mahdi, A.E. (2009), "An automated syllabus digital library system for higher education in Ireland", *The Electronic Library*, Vol. 27 No. 4, pp.640 - 658. ISSN: 0264-0473.

The next 250 years:
Ensuring long term learning with the British Library's Digital Library
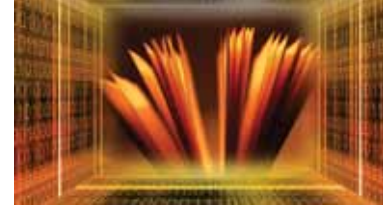
Introduction

The British Library's Digital Library Programme is now some seven years old. What are the goals of the programme? What priorities have been set? What mechanisms have been devised to provide access to digital material? How is the British Library going to ensure that digital resources are stored and made available in such a way that they can be permanently available? A case study by Steve Green throws light on these questions.

Because so much content is now available, it raises the question as to whether the British Library should be incurring heavy expenditure on building a digital store. The British Library decided that it should, firstly, because there is a need to ensure access for the long-term. The Library is in a very good position to do this, not least because it is a core activity of the Library, and deeply embedded in its culture. This is not true of commercial organizations, the needs of the business must take priority.

The second reason for the British Library taking on the preservation and accessibility role is that it already has a long history – 250 years and counting. It is committed to the long-term.

Goals and achievements

The goals of the Digital Library Programme are to:

• Receive, process, store and preserve all types of digital material "in perpetuity";

• Provide access to the material for those with the relevant right to use it;

• Ensure that material is easy to find;

• Ensure the authenticity of material;

• Ensure that the material can be consulted using contemporary applications;

• Ensure that patrons can experience the material with its original look and feel whenever possible.

The way in which the Digital Library Programme should achieve this have been defined as follows:

• Acquiring, buying or building software components to make up the Digital Library software, as necessary;

• Buying and deploying hardware;

• Defining procedures for handling digital content;

• Ensuring the Library's staff are properly trained;

• Defining standards in partnership with other libraries and relevant institutions nationally and internationally.

So far, the Programme has achieved a robust, secure infrastructure for the acquisition, processing, storage and discovery of, and access to, digital material. In March 2009, it contained over 440,000 items, including more than 60,000 nineteenth century books digitized from the Library's own collections and 11,000 sound files. Three copies of each item are stored: one at the site in Boston Spa in West Yorkshire, one at the London St Pancras site, and one at the National Library of Wales in Aberystwyth. All the materials can be accessed in the reading rooms at the St Pancras site.

Challenges

Despite that success, the Programme still faces challenges. Firstly, the British Library has the capacity to create and acquire much more than it can afford to store. Prioritization is necessary, but difficult to achieve because any items rejected may be lost forever.

Secondly, there is no widely adopted standard for publishing digital materials, which increase challenges, costs and time taken to process materials. Planning is a major challenge because of the scale of the project, and there are important decisions to take such as whether cataloguing is relevant in the age of full-text searching. The publishing industry itself is in a state of flux, with old business models becoming irrelevant. Many publishers are losing money and therefore there is an aversion to risk – and supplying content to the British Library reading rooms is seen as a risk.

Comment

This review is based on "The digital library programme at the British Library: goals and priorities" by Steve Green. Although there are undoubtedly significant challenges still, the British Library's digitization programme has achieved a great deal, and can provide useful lessons to others around the world who wish to ensure long-term preservation of and access to their cultural assets. The body of technical expertise which has been built up is itself of considerable value – to the Library, and to other institutions in the UK and overseas.

Reference

Green, S. (2009), "The digital library programme at the British Library: goals and priorities", *Interlending & Document Supply*, Vol. 37 No. 3, pp.136 - 139. ISSN: 0264-1615.

> " Publishers are making significant investments in ensuring that their content is held securely, digitizing back catalogues and making content available online (using both pay for and free of charge models). "