Library Hi Tech

Academic library and information services: new paradigms for the digital age. Proceedings of the 8th International Bielefeld Conference, 7-9 February 2006, Bielefeld, Germany

Guest Editors: Norbert Lossau and Sabine Rahmsdorf

www.emeraldinsight.com

Emerald

Library Hi Tech

Academic library and information services: new paradigms for the digital age. Proceedings of the 8th International Bielefeld Conference, 7-9 February 2006, Bielefeld, Germany

Guest Editors
Norbert Lossau and Sabine Rahmsdorf

CONTENTS

Access this journal electronically

The current and past volumes of this journal are available at:

www.emeraldinsight.com/0737-8831.htm

You can also search more than 100 additional Emerald journals in Emerald Fulltext (**www.emeraldinsight.com/ft**) and Emerald Management Xtra (**www.emeraldinsight.com/emx**)

See page following contents for full details of what your access includes.

Emerald

CONTENTS
continued

Access this journal online

As a subscriber to this journal, you can benefit from instant, electronic access to this title via Emerald Fulltext or Emerald Management Xtra. Your access includes a variety of features that increase the value of your journal subscription.

How to access this journal electronically

To benefit from electronic access to this journal, please contact **support@emeraldinsight.com** A set of login details will then be provided to you. Should you wish to access via IP, please provide these details in your e-mail. Once registration is completed, your institution will have instant access to all articles through the journal's Table of Contents page at **www.emeraldinsight.com/0737-8831.htm** More information about the journal is also available at **www.emeraldinsight.com/ lht.htm**

Our liberal institution-wide licence allows everyone within your institution to access your journal electronically, making your subscription more cost-effective. Our web site has been designed to provide you with a comprehensive, simple system that needs only minimum administration. Access is available via IP authentication or username and password.

Key features of Emerald electronic journals

Automatic permission to make up to 25 copies of individual articles
This facility can be used for training purposes, course notes, seminars etc. This only applies to articles of which Emerald owns copyright. For further details visit **www.emeraldinsight.com/ copyright**

Online publishing and archiving
As well as current volumes of the journal, you can also gain access to past volumes on the internet via Emerald Fulltext or Emerald Management Xtra. You can browse or search these databases for relevant articles.

Key readings
This feature provides abstracts of related articles chosen by the journal editor, selected to provide readers with current awareness of interesting articles from other publications in the field.

Reference linking
Direct links from the journal article references to abstracts of the most influential articles cited. Where possible, this link is to the full text of the article.

E-mail an article
Allows users to e-mail links to relevant and interesting articles to another computer for later use, reference or printing purposes.

Structured abstracts
Emerald structured abstracts provide consistent, clear and informative summaries of the content of the articles, allowing faster evaluation of papers.

Additional complimentary services available

Your access includes a variety of features that add to the functionality and value of your journal subscription:

E-mail alert services
These services allow you to be kept up to date with the latest additions to the journal via e-mail, as soon as new material enters the database. Further information about the services available can be found at **www.emeraldinsight.com/alerts**

Connections
An online meeting place for the research community where researchers present their own work and interests and seek other researchers for future projects. Register yourself or search our database of researchers at **www.emeraldinsight.com/ connections**

Emerald online training services
You can also access this journal online. Visit **www.emeraldinsight.com/training** and take an Emerald online tour to help you get the most from your subscription.

Choice of access

Electronic access to this journal is available via a number of channels. Our web site **www.emeraldinsight.com** is the recommended means of electronic access, as it provides fully searchable and value added access to the complete content of the journal. However, you can also access and search the article content of this journal through the following journal delivery services:

EBSCOHost Electronic Journals Service
ejournals.ebsco.com

Informatics J-Gate
www.j-gate.informindia.co.in

Ingenta
www.ingenta.com

Minerva Electronic Online Services
www.minerva.at

OCLC FirstSearch
www.oclc.org/firstsearch

SilverLinker
www.ovid.com

SwetsWise
www.swetswise.com

Emerald Customer Support

For customer support and technical help contact:
E-mail **support@emeraldinsight.com**
Web **www.emeraldinsight.com/customercharter**
Tel +44 (0) 1274 785278
Fax +44 (0) 1274 785204

Guest editorial

What are the new paradigms for academic libraries in the digital age? What are the new paradigms for information services in the digital age?

Let me briefly introduce you to the main topics of the 8th International Bielefeld Conference: topics, which can be expected to be of key relevance for the strategy of libraries in the digital age. You will find a representative sample of the conference papers in this volume. In addition, presentations can be found on the conference website (http://conference.ub.uni-bielefeld.de/2006/docs/).

477

Transnational networks and joint strategies

No one library in the world is today in a position to develop information services on their own.

The plethora of digital information, transnational search services and subject gateways, international digital libraries, the definition of information description and access standards, new models of scholarly publishing, and the preservation of cultural heritage and science records – we could easily continue to specify the apparently endless list of new competences, concepts, policies and technologies for libraries.

In the pre-digital world librarians were free to collaborate; in the digital, globally networked world of information, librarians have no other choice than working closely together.

In Germany, the Bertelsmann Foundation, together with the German Library Association (DBV), initiated a collaborative approach to formulate a joint strategy for German libraries by assembling a nationwide group of experts in a series of publications.

Four European countries (Denmark, Germany, the UK and The Netherlands) felt the need to support knowledge-sharing among libraries and information service providers. The result is the new Knowledge Exchange Office in Copenhagen. The German Goethe-Institute, which represents German culture abroad, has a well organised system in place, comprising a widely dispersed, decentralised network of libraries, supported by centres which provide a shared infrastructure for regions around the world.

Rethinking information services I – challenges and support for library managers

Where are we heading as academic information service providers? What will be our areas of activity, where the focus of our strategy? How can library managers cope with the new challenges for academic information services?

On the one hand, there is the need to justify the resources invested into our services. What is the balance between investment and outcome for the users? On the other hand, libraries need a much clearer picture of their users' needs. "Customer intelligence" is a marketing concept in the commercial world, which will find its way into library managers' strategy repertoire.

Are new converged organisational models for libraries and other academic service units the solution for integrated service concepts? Or can integrated services be equally offered by collaborating service units, without a tighter organisational umbrella?

Rethinking information services II – innovative services and future challenges for libraries

Offering support infrastructure for scholarly publishing and e-science will become, beyond repositories and preservation, two main areas of activity for academic libraries. E-science is a new approach which regards other types of information, in particular research data, as an integral part of a future information service infrastructure. But e-science is more than just research data: it encompasses new ways of generating knowledge – supported by information technology – within all academic disciplines and it considers a self-organising middleware infrastructure within scientific communities as an important element.

What will the role of libraries in e-science be? And how do digital libraries relate to e-science? The concept and technology of digital libraries itself will change in the future, and libraries and computer scientists will work more closely together.

Searching scholarly information in the digital age

Preparing the Bielefeld Conference, I came across a new book, written by John Battelle (2005), co-founding editor of *Wired* and founder of *The Industry Standard*, both journals closely linked to the new Internet economy. The book is entitled: *The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed our Culture*. That is it! The title of this book expresses explicitly one of the major paradigm changes for academic libraries and their services. The search for scholarly information, a key activity of libraries for centuries, has become a competitive market – and a prosperous business. According to John Regazzi (2004), Elsevier's Managing Director of Market Development and CEO of Elsevier Inc., internet search engines have reached sales of US$3 billion and a market capitalization and estimated value of nearly US$30 billion in their last five years of development. Google entered the stock market in Summer 2004. Since then their share value has quadrupled.

It does not go beyond the power of imagination to think of scholarly information discovery environments without our current online library catalogues and databases. Instead we may see scholarly search engines like Google Scholar, Scirus, Cite Seer, Microsoft and Yahoo – and others to come – as the search tools of choice for our users. I personally would not be afraid of these scenarios, as long as libraries can choose between a range of service providers and as long as we can integrate external services through standard technology, such as web services, into modernised library information service systems. Bielefeld University Library has been taking the first steps towards integrating Google Scholar and selected functionalities into BASE, the Bielefeld Academic Search Engine. From our point-of-view, it makes good sense to re-use existing services instead of re-building everything from scratch. But those external service providers must be willing to become part of other systems, and so far we have made good progress with Google Scholar.

Hands-on workshops

For the first time, the 8th International Bielefeld Conference organised "hands-on workshops", offering reports on practical implementations and concepts. The thematic areas covered three central topics for libraries in the digital age:

(1) National and international networking of digital repositories, which promises to form the new, cross-institutional infrastructure for all types of science and scholarship records. The relevance of this topic has just been underpinned by a European grant for Digital Repository Infrastructure Vision for European Research (DRIVER), a project that aims to build the testbed for a new repository infrastructure in Europe.

(2) Archiving and preserving the records of science through nationwide coordinated networks and projects.

(3) Marketing approaches for libraries, placing the users' interests and demands at the heart of library managers' strategies.

Conclusion and acknowledgements

The programme committee for the 8th International Bielefeld Conference assembled a programme, which proved to be the right framework to discuss many of the aspects described above. The impressive number of almost 500 participants from 35 countries, who again came together in Bielefeld, was encouraging for the organisers to continue their efforts with this conference. My warmest thanks go to my colleagues on the Programme Committee, Ronald Milne from Oxford University (UK) and Hans Geleijnse from Tilburg University (NL).

Summing up my thanks to everybody who helped to make this conference happen, I have to emphasise the unbroken enthusiasm of my own colleagues at Bielefeld University Library. Every second year, when the conference preparations reach their peak, almost 100 of my colleagues lend a hand to carry out all the upcoming multiple smaller and bigger tasks. This conference, I must stress again, would not exist without the unanimous support of all library staff.

My special thanks also goes to Emma Leeson, Oxford University Libraries, for her assistance in preparing the proceedings.

I wish you an inspiring and thought-provoking read.

Norbert Lossau
Bielefeld University, Germany

References

Battelle, J. (2005), *The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed our Culture*, Penguin, New York, NY.

Regazzi, J. (2004), "Beyond access and retrieval: the battle for mindshare", *Library Connect*, Vol. 2 No. 3, pp. 6-8.

8th International Bielefeld Conference, 7-9 February 2006, Bielefeld, Germany
Programme

Tuesday, 7 February 2006

- 13.00 *Welcome addresses*; *Dieter Timmermann*, Rector, Bielefeld University; *Ronald Milne*, Acting Director, Oxford University Library Services and Bodley's Librarian, UK; and *Norbert Lossau*, Library Director, CIO Scholarly Information, Bielefeld University.

- 13.30 *Keynote*. *Lynne Brindley*, Chief Executive, The British Library, UK, *Redefining the Library*.

Joint Strategies and Transnational Networks for Academic Libraries and Information Services. *Chair: Ronald Milne*, Acting Director, Oxford University Library Services and Bodley's Librarian, UK.

- 14.30 *Christian Hasiewicz*, Project Director, Bertelsmann Foundation. Towards a Future-safe Library – Current Trends and Developments in Germany (in German).

- 15.15 *Coffee break*.

- 15.45 *Sigrun Eckelmann*, Programme Director, Academic Libraries and Information Systems (LIS 4), German Research Foundation (DFG). Knowledge Exchange: Strategy and Roadmap for a New Network of Transnational Information Service, an Initiative of Four Partner Funding Organizations.

- 16.15 *Christel Mahnke*, Head of Section, Library and Information Services, Goethe-Institut München. WSIS and IFLA, UNESCO and GATS: Networks for Libraries on the International Level (in German).

- 16.45 *Roswitha Poll*, Former Chief Librarian, University and Regional Library Münster. *Philip Payne*, Librarian, Birkbeck College, University of London, UK. Impact Measures for Libraries and Information Services.

- 17.30 *Exhibition and Demonstrations in the Exhibition Showroom*.

- 19.00 *Conference Reception* by the Mayor of Bielefeld, Eberhard David, and the President of Bielefeld University, Dieter Timmermann.

- 20.30 *Speakers' Dinner (on separate invitation only)*.

Wednesday, 8 February 2006

Rethinking Information Services I – Challenges and Support for Library Managers. *Chair: Wolfram Neubauer*, Director, ETH-Bibliothek, Swiss Federal Institute of Technology Zurich, Switzerland.

- 9.00 *Michael A. Keller*, University Librarian, Director of Academic Information Resources, Publisher of HighWire Press, Publisher of the Stanford University

Press, Stanford University, US. Whither Academic Information Services in the Perfect Storm of the Early Twenty-first century?

- 9.30 *Reinhold Decker*, Professor of Marketing, Department of Business Administration and Economics, University of Bielefeld. *Michael Höppner*, Deputy Library Director, University of Bielefeld. Information Services: Customer Intelligence and Strategic Planning (in German).

- 10.00 *Mel Collier*, Library Director and Professor, Catholic University of Leuven, Belgium, Research Professor, Northumbria University, UK. Convergence of Libraries with Other Academic Services – Is It Relevant for European Universities? – Conclusions from a Recent Study.

- 10.45 *Coffee break*.

Hands-on Workshops, Concurrent Sessions (only in English). 11.15-13.00.
Session 1 – Archiving/E-depositing. *Chair: Elmar Mittler*, Director, Göttingen State and University Library.

- *Reinhard Altenhöner*, Head of IT, Die Deutsche Bibliothek, Frankfurt am Main Data for the Future: the German Project "Co-operative Development of a Long-term Digital Information Archive" (kopal).

- *Richard Boulderstone*, Director of e-Strategy and Information Systems, The British Library, UK. The British Library's Archiving Concept and Implementation.

- *Erik Oltmans*, Head Acquisitions & Processing Division, National Library of The Netherlands. The KB E-Depot: Permanent Access to the Records of Science.

Session 2 – Marketing Approaches in the Library Context. *Chair: Roswitha Poll*, Former Chief Librarian, University and Regional Library Münster.

- *Antonia Hermelbracht*, Academic Assistant at the Chair of Marketing, Department of Business Administration and Economics, University of Bielefeld. *Bettina Koeper*, Administrative Librarian, Bielefeld University Library. ProSeBiCA (Development of New Library Services by Means of Conjoint Analysis).

- *Sayeed Choudhury*, Associate Director for Library Digital Programs, The Sheridan Libraries of the Johns Hopkins University, US. CAPM (Comprehensive Access to Printed Materials).

- *Colleen Cook*, Dean, Texas A&M University Libraries, US. *Fred Heath*, Vice Provost and Director, University of Texas Libraries, US. LibQUAL + TM.

Session 3 – Networking Institutional Repositories. *Chair: Johannes Fournier*, Programme Director, Academic Libraries and Information Systems (LIS 3), German Research Foundation (DFG).

- *Leo Waaijers*, Platform Manager ICT and Research, SURF, The Netherlands. How do you DARE? Lessons learned from the DARE Programme in The Netherlands.

- *Mogens Sandfær*, Director of Center for Knowledge Technology, Technical University of Denmark. Networking Institutional Repositories in Denmark and Scandinavia.

- *Vanessa Proudman*, NEREUS Project Manager, University of Tilburg, The Netherlands. Nereus: One International Subject-based Repository Meeting Two Needs – Libraries and Researchers Collaborate.
- *Dirk Pieper*, Head of Acquisitions & Digital Library Coordinator. *Friedrich Summann*, Head of IT, Bielefeld University Library. Bielefeld Academic Search Engine (BASE): an End-user Oriented Institutional Repository Search Service.
- *Susanne Dobratz*, Head Joint Electronic Publishing Group of University Library and Computer and Media Services, Humboldt-University of Berlin. DINI Institutional Repository Certification and Beyond.
- 13.00-15.00 *Lunch and exhibition*.

Rethinking Information Services II – Innovative Services and Future Challenges for Libraries. *Chair: Michael A. Keller*, University Librarian, Director of Academic Information Resources, Publisher of HighWire Press, Publisher of the Stanford University Press, Stanford University, US.

- 15.00 *Sarah E. Thomas*, University Librarian, Cornell University Library, US. Publishing Solutions for Contemporary Scholars: The Library as Innovator and Partner.
- 15.30 *Donatella Castelli*, Scientific Coordinator of DILIGENT, CNR-ISTI, Italy. Digital Libraries of the Future – and the Role of Libraries.
- 16.00 *Liz Lyon*, Director, UKOLN, University of Bath, UK. E-Science and the Role of Libraries.
- 16.30 *Tony Hey*, Corporate Vice President of Technical Computing, Microsoft Inc., US. E-Science and its Implications for the Library Community.
- 17.15 *Exhibition and Demonstrations in the Exhibition Showroom*.
- 19.30 *Conference Dinner at the Mövenpick Hotel*.

Thursday, 9 February 2006
Searching Scholarly Information in the Digital Age: Challenges and Opportunities for Academic Libraries. *Chair: Jürgen Bunzel*, Programme Director, Academic Libraries and Information Systems (LIS 1), German Research Foundation (DFG).

- 9.00 *Arie Jongejan*, CEO, Swets Information Services, The Netherlands. Scholarly Information on the Web – a Vision for the Future Service Landscape.
- 9.30 *Wim Jansen*, Scientific Officer of the European Commission, DG INFSO, European Commission, Belgium. E-Infrastructure for the European Research Area – a Roadmap.
- 10.00 *Dirk Lewandowski*, Project Manager, Search Engine Quality, Department of Information Science, Heinrich-Heine-University Düsseldorf. Exploring the Academic Invisible Web (in German).
- 10.45 Coffee break.

Searching Scholarly Information in the Digital Age: Challenges and Opportunities for Academic Libraries. *Chair: Norbert Lossau*, Library Director, CIO Scholarly Information, Bielefeld University.

- 11.15 *Anurag Acharya*, Principal Engineer, Google Inc., US. Searching Scholarly Literature: A Google Scholar Perspective.
- 11.45 *Bjørn Olstad*, CTO, Fast Search & Transfer ASA (FAST), Norway. How Emerging Technology Will Drive New Precision Levels for Academic Libraries.
- 12.15 *Ammy Vogtlander*, Director Search, Elsevier, The Netherlands. Finding Scholarly Content in the Age of Web Search.
- 13.00 *Hans Geleijnse*, Director Library/IT-Services, CIO, Tilburg University, The Netherlands. Wrap-Up of the Conference and Farewell Address.
- 13.15-14.00 *Farewell snack*

Programme committee
Norbert Lossau, Library Director, CIO Scholarly Information, Bielefeld University. *Ronald Milne*, Acting Director, Oxford University Library Services and Bodley's Librarian, UK. *Hans Geleijnse*, Director Library/IT-Services, CIO, Tilburg University, The Netherlands.

Re-defining the library

Lynne Brindley
The British Library, London, UK

Abstract

Purpose – This article was originally the keynote speech to the Bielefeld Conference. The article aims to explore the challenges facing libraries in the digital age and consider ways in which they need to reshape and rethink their services and skills to maintain their relevance and contribution.

Design/methodology/approach – Provides a review of a wide range of recently published materials (2003-2006) and gives a broad perspective on the challenges facing libraries. These are then considered within the case study experience of the British Library to identify key themes for redefining the concept of "library".

Findings – This article gives a clear articulation of the challenges facing libraries. Through the case study the author identifies seven themes as central to redefining the library in the twenty-first century: know your users and keep close to them; re-think the physical spaces and create a desirable draw; integrate marketing into the organisation; open up legacy print collections to digital channels and through digitisation; reduce legacy costs and continue to improve productivity in traditional activities; invest more in innovation and digital activities; and develop our people and ensure the right mix of skills.

Practical implications – A practical source of ideas for those seeking to develop their own library activities and a thought-provoking analysis for anyone interested in the implications of the digital age.

Originality/value – This paper gives an original view of changes within the library sector from one of the leaders in the field and is rooted in the practical and innovative approaches adopted by one of the world's great research libraries.

Keywords Libraries, Change management, Strategic management

Paper type Conceptual paper

I am delighted and honoured to be giving the keynote speech at this important and prestigious conference of experts. Thank you for the invitation. There is one advantage of giving a keynote speech, namely that no-one has already covered what you are about to say (as so frequently happens when you are later on the programme) and it is the prerogative of the keynote speaker to ask more questions than to give answers! I am sure that I will be doing just that, for the theme of the conference – re-defining, re-thinking, developing new paradigms for library and information services in the twenty-first century – is fundamental to all of us and to the future health of scholarship and research in all our countries and in our institutions and for those we serve. This is a formidable leadership responsibility.

On a less serious note I have also decided to spare you a power-point presentation. I am sure there will be plenty of those to come: indeed it is my normally preferred style but I think that a keynote speech should be just that!

Flavour of the challenges

When I was starting to prepare for this paper – blank screen in front of me, several cups of coffee drunk and all possible displacement activities completed – I happened to receive the January 2006 copy of *Information World Review* (2006) through the post at

home. I needed to look no further than scanning the topics of the first few pages to indicate a flavour of the challenges we face. Let me share a few with you:

- the accuracy of Wikipedia articles on science is validated by Nature. The error level is not significantly greater than levels in Encyclopaedia Britannica;

- libraries are urged to embrace the second generation of Internet technology Web 2.0 to satisfy user demands, saying that existing library catalogue standards, such as MARC and Z39.50 need to be replaced by XML technology, enabling access to information from a wide variety of web services; library catalogues are compared unfavourably to Amazon and Google search services;

- the now rather done to death topic of open access is covered with a suggestion of a split between science, technology and medicine publishers, Oxford University Press, Blackwell Publishing and Springer all of whom have signed up to the Wellcome Trust Open Access model; and the Dutch (Reed Elsevier, Wolters Kluwer) and the North Americans (Wiley, Sage) who appear to be steadfastly unmoved by open access;

- Michael Gorman slams digitisation of scholarly texts as a waste of money while Elisabeth Niggemann, Director of Die Deutsche Bibliothek urges both public (European Union) and private sector investment to ensure wider access to collections; and

- Apple's flagship store in Regent Street in London is projected as a model for libraries as they reshape space and service provision to encourage knowledge exchange and solving information problems.

There is also a review of the year (Chillingworth, 2006) – *The Stories that Rocked your World in 2005* – and just a sample of topics is indicative of the complexity and rapidity of change within our information world over just 12 months.

In January 2005 Google announced its plans to digitise the collections of five major libraries, including the Bodleian at Oxford. Later initiatives include the Open Content Alliance (OCA) with Yahoo and Microsoft; the British Library/Microsoft agreement to digitise millions of pages; Macmillans joined the digitisation goldrush with BookStore; publishers and authors hit Google with lawsuits around copyright violation, and Amazon announced a new service to change the model to purchase book content by the page or chapter. Digitisation has never been such a hot topic! Content may really be king, at least for a little while!

But in the UK subject librarians were being threatened with redundancy at Bangor University; they were not considered as offering value for money compared with the net; and at the Science Museum the library collections were to be split up in the face of a financial crisis.

Regulatory and legislative issues pertaining to the digital environment surfaced strongly. In the UK the Freedom of Information Act came into force; in a year when terrorist bombs struck in London, terrorism was never far from the headlines with new legislation on police powers and criminal evidence and a new terrorism bill all threatening to compromise librarians and their normal business activities. And as the year ended the UK government announced a fundamental review of intellectual property in the digital environment to take place in 2006.

What does all this mean?

I have laboured these developments, to some extent a rather random sample, because they seem to me to represent either explicitly or implicitly the nature of the serious challenge to libraries and to information professionals in this first decade of the twenty-first century. They suggest a picture of ever more rapid innovation, mostly happening outside libraries and driven from the commercial sector; a picture of confusion and contradiction in the range of business models that are emerging and being experimented with; and of new demands from discerning and empowered users. Let me quote from David Warlock (cited in *Information World Review*, 2006) a well-respected industry watcher who says:

> […] users are in control on the networks, and exhibit a power unparalleled in the legacy world of print […] the gold […] is the greater ability of users to discriminate, select, personalise and customise, as powerful players in the information industry network (p. 32).

But such a challenge is also an exciting opportunity for the library and information sector to play new roles and to define a new future in a very fast moving and competitive environment. There is in any case no choice but to change, and change quickly if we wish to remain relevant for the future.

Challenges for the library and information sector

The challenge for libraries in the twenty-first century, as now only one part of a great diversity of alternatives, is to find new ways to add value and remain relevant in this rapidly changing, confusing and competitive environment. While the distant future for libraries is not clear, it is timely for libraries to challenge some historic assumptions and ask some fundamental strategic questions:

- How can we serve the needs of the digitally savvy, impatient Google generation for whom the web – a global information commons – has primacy of place for information and knowledge seeking?

- How can we continue to enable the research and learning process when increasingly it is happening in a virtual realm outside the context of the library?

- How can we be relevant to those who have never set foot in a library – to provide the infinite connectivity to information with its stacks in the ether?

- Does the library as place have relevance and how should space be best used?

- Where should we focus in the information value chain, and what should we not do?

- How can libraries provide effective stewardship of both digital and physical collections, and what is our role regarding non-traditional information types, such as e-science data?

- How are publishing and intellectual property regimes changing, and how must we influence thinking on them and change in response?

- What types of skills do libraries need to exploit advances in technology and informatics, both to enhance knowledge exploration and presentation and to enable new ways of searching and mining their collections?

- What types of collaboration and alliances do libraries need to engage in to present coherent collections and to create innovative new products and services for content delivery?

These were some of the questions we asked ourselves at the British Library when we were formulating our new strategy in 2004/2005 (The British Library, 2005). Technology is arguably turning on its head our assumptions about our value, it is challenging the roles of all accepted players; and it is enabling increasingly promiscuous users with different and higher needs to have much wider choice to fit their digital lifestyles.

Perceptions of libraries and information resources

I hope that I have made at least the outline case for major, transformational change in the library and information sector, driven from the imperatives of the external environment – the information industry, the technology and most importantly the demands of the users.

I want now to move on to some observations of how libraries are perceived by their users, to give context to the nature of the change and re-definition that may be required. In this I have been greatly assisted by a recent report published by the Online Computer Library Center (2005). If you have not already seen the report I commend it warmly to you as a rich, international source of market research data on perceptions of the library and its services, which should act as another wake-up call for us all.

The report followed on from an earlier environmental scan identifying some dissonance in expectations of libraries. User priorities were seen as ease of use, convenience and availability, all regarded as equally important to the information consumer as information quality and trustworthiness (Online Computer Library Center, 2003a). The objectives of the 2005 study were to ascertain how libraries are perceived by today's information consumer and to see whether libraries still matter and what future patterns of library use might be. With this acknowledgement let me pick out some of the salient headlines.

Libraries are seen as more trustworthy/credible and as providing more accurate information than search engines. Search engines are seen as more reliable, cost-effective, easy to use, convenient and fast. The library is not the first or only stop for many information seekers. Search engines are the favourite place to begin a search and respondents indicate that Google is the search engine most recently used to begin their searches. In addition users wanted "more books" and longer and more convenient opening hours. Perhaps no surprises there!

Through increasing familiarity with search engines and the web comes greater self-reliance of information consumers, who feel confident in their own evaluation of sources of all kinds. Survey respondents are generally satisfied with libraries and librarians but most do not plan to increase their use of libraries. Indeed the brand association of libraries appears to be rather depressingly nostalgic, traditional, and focussed on books. Even with their strong emotional attachment to the idea of the library there was clear dissatisfaction with the physical and service experience of the libraries they use. Poor signage, inhospitable surroundings, unfriendly staff, lack of parking, dirt, cold, hard-to-use systems and inconvenient hours were repeatedly mentioned.

And finally more on brand image. Most respondents feel that library is synonymous with books. Books dominated responses across all regions surveyed and across all age groups, despite libraries' growing investment in electronic resources and digital activities.

In summary, these findings do not make comfortable reading. I am sure that each one of you will wish to argue that the findings do not apply to your particular institution or

type of library, but that might suggest a level of denial that would be inappropriate. The opportunities are there for significant re-branding and re-positioning, both in terms of the design and delivery of digital and physical services, recognising that the information landscape will if anything become even more competitive and consumers will become even more discerning and willing to take what information they need from wherever they can most conveniently and painlessly get it. The call to action suggests the need for much deeper understanding of each of our user communities; a much more developed sense of place as social context for services; and greater attention to relevance, distinctiveness and convenience of all of our future services.

How libraries stack up

Lest we get too concerned I would draw your attention to yet another Online Computer Library Center (2003b) report that through a range of statistical data re-affirms the importance and comparative scale of library activity, primarily but not exclusively in a US context. The basis on which new opportunities can be built is immense:

- US public library cardholders outnumber Amazon customers by almost 5 to 1;
- each day, US libraries circulate nearly four times more items than Amazon handles;
- one out of every six people in the world is a registered library user;
- five times more people visit US public libraries each year than attend US professional and college football, basketball, baseball and hockey games combined;
- there are over 1 million libraries worldwide with 16 billion volumes; and
- there are some 690,000 librarians worldwide.

We already know how well information professionals network (virtually and physically); we already know how willing we are to share experience and best practice. These are strong characteristics that should serve us well as we seek to ramp up the scale and pace of change.

How do other industries respond to change?

I would like to turn now briefly to what, if any, help and understanding we might get from models for change more generally. When considering the issue of re-definition in the context of the British Library I have found the framework offered by McGahan (2004) in a *Harvard Business Review* article entitled "How industries change". She argues strongly that to develop strategy and make appropriate investments in innovation within the organisation requires a real understanding of the nature of change within the industry. This is of course easy to say but difficult to assess, particularly to take a long-term look in a rapidly changing short-term context. The business world is littered with misinterpretations of signs. She suggests, however, four distinctive trajectories of change – radical, progressive, creative, and intermediating.

I believe that libraries and information services (depending on their particular nature) are operating in an environment between intermediating and radical change. An industry on a radical change trajectory is entirely transformed, probably over a timescale of decades, with an end result of complete reconfiguration (usually diminished). Companies dealing with radical transformation it is suggested should move strongly to improve productivity in existing activities without significant

investment, conduct experiments with new products and services and develop new distribution channels.

Intermediating change is more common than radical change. It is where the core assets – knowledge, brand, content, patents – retain much of their value if they are used in new ways. This requires the simultaneous preservation of valuable assets and re-structuring of key relationships, and means finding innovative and unconventional ways of extracting value from core resources. Managing this dual track approach is extremely challenging.

From a British Library perspective we continue to focus on increasing productivity and streamlining traditional processes (largely through systems changes); we are finding innovative ways of exploiting our core assets of content combined with expertise; and we are opening up new channels of delivery largely through digital partnerships and new service developments. This re-positioning in the digital library world, at the same time as sustaining our core statutory functions both for the print and digital domains, is a major leadership and management challenge requiring changes in structure, skills and investment patterns.

Digital library challenges

It is interesting to also consider the views of leading experts in digital library developments and associated research. Cliff Lynch (2005) in assessing prospects for digital libraries in the next decade suggests strongly that the major challenge is to:

Connect and integrate digital libraries with broader individual, group and societal activities, and doing this across meaningful time horizons that recognize digital libraries and related constructs as an integral and permanent part of the evolving information environment.

Additionally he argues that:

The issue of the future of libraries as social, cultural and community institutions, along with related questions about the character and treatment of what we have come to call "intellectual property" in our society, form perhaps the most central of the core questions within the discipline of digital libraries – and that these questions are too important to be left to libraries, who should be seen as nothing more than one group among a broad array of stakeholders.

This questioning of "what is a digital library anymore, anyway" is echoed in a challenging article of this title by Carl Lagoze *et al.* (2005) at Cornell University. He worries about a perception that the "googlization" of digital libraries and information more generally means that digital library problems have either already been solved or will be solved by Google, MSN, Yahoo! and others. One might regard this as simply a plea for more research funding from interested parties, or perhaps more seriously as a need for us to think well beyond search and access as presently conceived towards a much richer information environment for information sharing, aggregation, manipulation, collaborative working, and indeed digital preservation which Cliff Lynch sees as an enormous, fundamental societal issue for the next decade.

So where is all this leading us and what should we be doing?

I hope that so far in this keynote speech I have painted a broad picture of an increasingly challenging information environment within which libraries of all types operate. Just reflecting on developments within the last twelve months the pace of

change is probably faster than any of us has experienced in our professional lives so far. In addition there is much evidence that in this new context whilst libraries are still valued there are increasing signs of dissatisfaction and perceptions of a lessening value compared with other options. However, the changing environment provides opportunities for re-defining our roles and our relevance and for developing new roles in a much wider range of public and private partnerships and collaborations. There is a need for strong strategic and thought leadership, more risk-taking than has generally been associated with libraries, and greater understanding of and attention to the changing demands of our different user populations.

I would like therefore in the final part of my talk to pull together some themes of general strategic relevance for the conference, drawing on our experience of major change, re-branding and strategy development at the British Library over the past few years. The British Library has become more externally-focussed and market-facing; we have rationalised and modernised our portfolio of services and have a higher public and government profile. We have faced up to difficult staffing issues, have brought in people with new and complementary skills and have invested to catch up on our technology infrastructure and do leading work in areas such as digital preservation.

Re-defining the library
I would like now to offer some themes that I believe will be central to the continuing re-definition and re-positioning of libraries to remain relevant in the twenty-first century. They are likely to be more or less relevant to you depending on the context of your service and are certainly not comprehensive.

1. Know your users and keep close to them (and your lost users and your non-users)
If there is a common message coming across from all those thinking deeply about the future of libraries it is that we need to be more deeply involved with our users to really understand how their work patterns are changing, to anticipate their future requirements and how information services can be better integrated with the increasingly digital life-style of new generations of students, researchers and knowledge workers.

The British Library had, like many libraries, for a long time viewed its users as a homogenous group of "readers". In 2000, we looked closely at this group and identified, unsurprisingly, that all readers were not the same. Using expertise from commercial marketing we identified clear audiences, all of whom had specific ideas of what they needed from the library.

Researchers, including staff in higher and further education; postgraduates; high R&D industries; writers and scholars; government researchers; the library network; schools and young people; individuals – including undergraduates – pursuing their own research projects; the wider, general public; and business, which was identified as a core audience which at that stage was under-utilising the wealth of resources the library had to offer.

Our core functions as a great research library are primary but our users and their needs are varied and distinct. Defining a library that will balance the needs of the humanities professor and the new entrepreneur and the undergraduate with approaching finals and the office worker looking for somewhere pleasant to spend their lunch break, is a challenge, but our new, focused understanding of who our users are is an important catalyst for creativity in approaching the rest of my core themes.

2. Re-think the physical spaces of the library and create a "desirable draw"
Libraries should aim to be uplifting, innovative and inspiring cultural, social and intellectual spaces, encouraging debate and collaboration, and desirable as places to be in, even in the age of ubiquitous internet access. Some of the best models for the future will come, not simply from existing libraries (although there are some stunning international examples) but also from bookshops, Apple stores, museums and galleries and new concepts in retailing spaces. If the British Library is the physical representation of the information age, what should it be like? Can the surroundings and atmosphere add value – even enhance – the information we are providing?

While defining our users, we identified that we were under-utilised by business and entrepreneurs: an important audience for ourselves and for our role as support to the UK's creative and technological economy. Market research showed that for every £1 of public investment we received, we generated £4.40 of national value. How could we expand the use business made of our resources? What could we do to make ourselves more appealing to them?

We found a good model in the New York Public Library's Science, Industry and Business Library and I am proud that, one month tomorrow, the British Library's Business and IP centre will open. The aim will be to deliver a coherent programme of support at each stage of the innovation lifecycle: information support and networking space, with expert support from partner organisations on business planning, financial development and intellectual property.

But rethinking our physical space need not always mean rebuilding it. The British Library's St Pancras site had just opened when I arrived in 2000, after a costly and protracted building process. Much of what we have done in the last five years has been about encouraging new audiences to come and use us as a public amenity. Every summer now we have live music sessions on the piazza and tying into our launch of Mozart's digitised musical diary our central foyer was filled by people listening to live performances of Mozart string quartets. We have installed London's largest public wi-fi space and are becoming a place of choice for laptop users as well as delivering an additional resource for our core researchers. And before Christmas we gained much media coverage for our Nobel Prize Exhibition through our second "Mingles" evening billed under the strap-line "In love with science and want to share your passion? Come and meet other like-minded single people and attend an open evening at the exhibition". Libraries can be sexy – over 400 new visitors came along that night.

3. Integrate marketing in your organisation and in the way you approach strategy and service development
My two previous themes have not been about changing the core library function of repository and steward, but have been about changing our attitudes to that function to take a proactive approach to engaging our users and marketing our resources to meet their needs. As we seek to redefine ourselves it is essential that we integrate marketing into the way we approach strategy and service development.

In 2000, for the first time we introduced a directorate for strategic marketing and communication, headed up by a respected commercial specialist to help us do this. Our marketing had traditionally concentrated in promotional marketing activity without sufficient emphasis on positioning or branding. Curators were the driving force behind much of what we did with the effect that too often we could be focused on our own areas

of interest and expertise rather than connected to those of our users. The new marketing director led on the tough questions – who are our users? What is our proposition to them? – and sought to engage all library staff in considering the answers.

The result has not been simply rethinking how we use our space, or improving signage and welcome, or remodelling our Treasures gallery to best display the jewels of the nation's heritage: rather, a dynamic synergy between curatorial and marketing staff has emerged. Exhibitions are now agreed and taken forward on the basis of whether they will meet a market need and then developed in tandem with promotional events. Our most successful exhibition drew upon curatorial expertise and knowledge and enhanced this with a marketing edge. The result was a journey down the Silk Road, including letters, maps, objects, the Diamond Sutra, the world's oldest printed book and drew record breaking numbers of visitors. It was supported by a series of seminars and lectures drawing in new audiences to hear a range of speakers and experts.

4. Open up your legacy print collections to digital channels and reveal them through digitisation

Extracting new value from core assets is a key part of managing "intermediating change". Many of the great libraries represented here have unique assets, particularly in their primary collections of archives, manuscripts and rare materials. In addition the scale and scope of secondary source material (even just considering the corpus of out of copyright material) that can be digitised and therefore made infinitely more visible and accessible to the world is enormous. There are great opportunities here for libraries to find new channels, deliver public value and ensure business models that enable sustainability. The latter, however, depends on some careful thinking on the value of the intellectual property over which we have stewardship balanced against many of our professional aspirations to open our content freely to everyone. It is always worth reflecting that initial one-off costs of digitisation pale into insignificance set against a commitment to perpetual access and preservation of the newly created digital asset.

The British Library's award winning Turning the Pages (TTP) programme and technology has been a direct product of our thinking about how digital can help us open up access to some of our most valuable items – and it is a source of immense pleasure to me that it has been seized by many of the institutions here today for doing that with their own treasures.

At its start TTP was about helping us give members of the public access to precious books while keeping the originals safely under glass, allowing them to virtually "turn" the pages of manuscripts in a realistic way, using touch-screen technology and interactive animation. But that initial step was just the beginning of the potential digitisation affords us as we rise to the challenges of "intermediating change".

The British Library has developed a set of assessment criteria to help us prioritise areas of our collections for digitisation. High on our list is our vast collection of newspapers, due to its fragile nature. Not only is there a huge collection of at risk material that could be digitised, but also, the time-based nature of newspapers makes them ideal to "map" and link to other historical collections and items. We are also exploring opportunities for digitising our unique audio collection. Audio resources are generally harder to manage and the resultant resource is more difficult to search, relying as it does on metadata. However much progress has been made in speech

recognition technology and automated full-text creation of audio content may soon be possible, making searching of speech based items much easier.

Digitisation also makes exciting reunification projects possible, increasing the world's knowledge and advancing digital scholarship. Three weeks ago we launched a virtually-united leaf of Mozart's musical diary – cut into two pieces 200 years ago by the widowed and impoverished Mrs Mozart to increase the total sale value. TTP has given the world access to previously unknown music and a catalogue of complete works integrating sound from the sound archive into the digitised work. The BBC covered the story and our website recorded a record number of hits as people linked through to us on the back of media coverage.

And many of you will already know about the unprecedented collaboration underway to digitally reunify the Codex Siniaticus involving all four of the institutions at which parts of the manuscript are held: St Catherine's Monastery, Sinai, the British Library, the University of Leipzig and the National Library of Russia.

Digitisation opens doors to new and dynamic partnerships. Last autumn the British Library announced its intention to work with Microsoft to digitise 100,000 out of copyright books and make them available over the internet. There are complex intellectual property issues involved in such partnership working, but I view the Microsoft deal as an example of how libraries can work with the new players in the information arena as we modernise and update our services.

5. Reduce legacy costs and continue to improve productivity in traditional library activities

Underpinning the functioning of our libraries are many traditional processes and activities – selection, acquisition, cataloguing, fetching and retrieving, preserving, and so on. We need to be vigilant to ensure that these well-worn routines continue to be challenged both in the way we do them and the priority we give them. Freeing up resources for investment in new things, and in that I include freeing up our best staff, as well as creating funds for research and technological innovation will be critical if we are to keep up with changing expectations. This sometimes means challenging long-assumed professional roles and other entrenched working arrangements.

For over 40 years the British Library's document supply service has fulfilled a critical role in the UK's information provision and remains the world's largest document supplier. At its peak level of demand in 1998/1999 – only six years ago – it was fulfilling over 4 million requests for individual documents annually – three million to UK customers and one million overseas. But demand has slowed in recent years because of the new opportunities provided for through the "big deals" and easier digital access.

Three years ago, the library embarked upon a programme of extensive modernisation of our document supply operation, partly funded by the UK government's Invest To Save Budget. The library replaced all photocopiers with state of the art scanning stations, which enable us to deliver all documents electronically regardless of their original format and which have reduced standard turnaround times from days to just hours.

6. Invest more in innovation and digital activities

One of our real challenges is to create enough resources for faster innovation and investment in experiments and new digital services. All of us have opportunities for

adding value to our communities through new roles such as institutional repository management, digital asset management and audit, digital scholarship, e-learning activities, and so on. Some of these roles require new kinds of consortial and other partnerships across the public and the private sector.

The Legal Deposit Libraries Act 2003 has given us responsibility for archiving the UK's digital output and we are preparing the infrastructure and methodology for this now. Nobody has exactly the product we are looking for, so we are using a combination approach, of buying in components and working with the supplier to develop them to meet our needs as DOM – our Digital Object Management System.

The first practical release of the storage system provides a preservation-quality digital store for material received under the voluntary deposit scheme. Subsequent releases will add more functionality to this. We will then extend this service to handle other types of material, initially e-journals. Other high priority materials are CDs, DVDs, and other "hand-held" items, and the growing number of digital newspapers. Technical direction of the solution architecture has been validated by two external technical advisory panels.

We are now focusing on ingest and are in the process of selecting a vendor. Requirements for other functions (data management, administration, and access) are being developed. Digital preservation is being addressed by a newly-recruited team and we are building European partnerships.

7. Develop our people and ensure that we have the right mix of skills
I believe that the role of librarian and information professional has never been more important, but what we mean by librarian has to be reconsidered. It is we, not just our institutions, who must face up to intermediating change.

The information professionals of the future need to be outward-going people, with really sharp business skills and a huge understanding of technology and the implications of the internet. They need to be able to understand and engage with users to bring their collections to life, in a way that a search engine on someone's desk simply cannot. They need to concentrate on what they are good at – information management, metadata, reference services, to name but a few – and be ruthless about bringing in specialisms they need from outside to add value to their core tasks.

Because it is the mixture that is so powerful. Very few of the examples I have given today of how we at the British Library are redefining ourselves for the twenty-first century would have been possible without the new blend of skills and expertise we have acquired over the last five years.

My final point is to encourage us all to have a much stronger voice in the grand challenge debates that are part of the development of the digital society and economy. There are many big and complex policy issues that are live at national government, European and indeed global level. As library and information professionals we should speak robustly to ensure appropriate balances between public interest and commercial imperatives in digital copyright and intellectual property regulation; we should be central in working on and piloting new and sustainable business models for digital services; and we should widen understanding and debate about the importance to society and to individuals of digital preservation.

Concluding remarks

I would like to conclude simply with a quote from the library's founders when they brought together the original collections in 1753 to be "preserved therein for public use, to all posterity" providing access to the world's knowledge for "all studious and curious persons".

At this early stage of the twenty-first century it is arguably the technological geeks who, in their fight to win a browser war, are doing most to re-interpret and fulfil the utopian dream of universal access. It is imperative that we, as the custodians of the world's knowledge create our own vision and contribution to this desired future by re-defining the library to be relevant for this and future generations.

References

(The) British Library (2005), *Redefining the Library: The British Library's Strategy 2005-2008*, The British Library, London, available at: www.bl.uk/about/strategy.html (accessed 24 April 2006).

Chillingworth, M. (2006), "The stories that rocked your world in 2005", *Information World Review*, Vol. 220, pp. 9-10.

Information World Review (2006), "Flashback", *Information World Review*, Vol. 220, p. 32.

Lagoze, C., Krafft, D.B., Payette, S. and Jesuroga, S. (2005), "What is a digital library anymore, anyway?", *D-Lib Magazine*, Vol. 11 No. 11, available at: www.dlib.org/dlib/november05/lagoze/11lagoze.html (accessed 24 April 2006).

Lynch, C. (2005), "Where do we go from here? The next decade for digital libraries", *D-Lib Magazine*, Vol. 11 Nos 7/8, available at: www.dlib.org/dlib/july05/lynch/07lynch (accessed 24 April 2006).

McGahan, A.M. (2004), "How industries change", *Harvard Business Review*, Vol. 82 No. 10, pp. 87-94.

Online Computer Library Center (OCLC) (2003a), *OCLC Environmental Scan: Pattern Recognition*, OCLC Inc., Dublin, OH, available at: www.oclc.org/reports/2003escan.htm (accessed 24 April 2006).

Online Computer Library Center (OCLC) (2003b), *Libraries: How They Stack Up*, OCLC Inc., Dublin, OH, available at: www.oclc.org/reports/2003libsstackup.htm (accessed 24 April 2006).

Online Computer Library Center (OCLC) (2005), *Perceptions of Libraries and Information Resources: A Report to the OCLC Membership*, OCLC Inc., Dublin, OH, available at: www.oclc.org/reports/2005perceptions.htm (accessed 24 April 2006).

Corresponding author

Lynne Brindley can be contacted at: chief-executive@bl.uk

Digital libraries of the future – and the role of libraries

Donatella Castelli
ISTI-CNR, Pisa, Italy

Abstract

Purpose – The purpose of this article is to introduce the digital libraries of the future, their enabling technologies and their organisational models.

Design/methodology/approach – The paper first discusses the requirements for the digital libraries of the future, then presents the DILIGENT infrastructure as a technological response to these requirements and, finally, it discusses the role that libraries can play in the organisational framework envisioned by DILIGENT.

Findings – Digital libraries of the future will give access to a large variety of multimedia and multi-type documents created by integrating content from many different heterogeneous sources that range from repositories of text, images, and audio-video, to scientific data archives, and databases. The digital library will provide a seamless environment where the co-operative access, filtering, manipulation, generation, and preservation of these documents will be supported as a continuous cycle. Users of the library will be both consumers and producers of information, either by themselves or in collaborations with other users. Policy ensuring mechanisms will guarantee that the information produced is visible only to those who have the appropriate rights to access it. The realisation of these new digital libraries requires both the provision of a new technology and a change in the role played by the libraries in the information access-production cycle.

Practical implications – Digital libraries of the future will be core instruments for serving a large class of applications, especially in the research field.

Originality/value – The paper briefly introduces one of the most innovative technologies for digital libraries, and it discusses how it contributes to the realisation of a novel digital libraries scenario.

Keywords Digital libraries, Knowledge management

Paper type Viewpoint

Introduction

Research on digital library (DL) systems started in Europe in the mid-1990s. At that time DLs were seen essentially as repositories of digital texts accessible through a search service that was operating by indexing information stored in a centralised metadata catalogue. The construction of a DL was very resource-consuming since, for each new DL, both the content and the software providing the DL functionality were built from scratch. As a result of this development approach, only powerful user communities[1] or user communities with in-house computer science technical skills (Leiner, 1998) could afford the building up of DLs. These DLs were created to serve end-users only as consumers of information. They did not provide any functionality for submitting the documents. The submission was usually performed either by the author or by a librarian operator by means of specific procedures residing outside the DL.

Today, the requirements imposed on DLs are very different from that early time. A novel notion of DLs, also referred to as "knowledge commons" (Ioannidis, 2005), has recently emerged, whose fulfilment requires new technologies and new organisational models. This paper focuses on such new DLs by first discussing the motivations for

their introduction, then presenting an innovative DL technology, called DILIGENT, and, finally, illustrating the role that libraries can play in this new scenario.

Digital libraries of the future

According to the most recent understanding, the DLs of the future will be able to operate over a large variety of information object types – far wider than those maintained today in physical libraries and archives. These information objects will be composed of several multi-type and multimedia components aggregated in an unlimited number of formats. These, for example, can mix text, tables of scientific data and images obtained by processing earth observation data, or they can integrate 3D images, annotations and videos. These new information objects will offer innovative and more powerful means to researchers for sharing and discussing the results of their work. In order to be able to support these objects, the DL functionality has to be appropriately extended far beyond that required to manipulate the simple digital surrogates of the physical objects. In order to support these objects the DL may need considerable resources. For example, the creation and handling of the new documents may require access to many different, large, heterogeneous information sources, the use of specialised services that process the objects stored in these sources for producing new information, and the exploitation of large processing capabilities for performing this tasks.

New DLs are also required to offer a much richer set of services to their users than in the past. In particular, they must support the activities of their users by providing functionalities that may range from general utilities, like annotation, summarisation or co-operative work support, to very audience-specific functions, like map processing, semantic analysis of images, or simulation. The availability of this new DL functionality can, in principle, change the way in which research is conducted. By exploiting such types of DL, for example, a scientist can annotate the article of a colleague with a programme that extracts useful information from a large amount of data collected by a specific scientific observatory. This programme, executed on demand when the annotation is accessed, can complement the content of the paper with continuously refreshed information.

In the new DLs users are not only consumers but also producers of information. By elaborating information gathered through the DL they can create new information objects that are published in the DL, thus enriching its content. The new DLs are thus required to offer services that support the authoring of these new objects and the workflows that lead to their publication.

In parallel with the above evolution of the role of DL systems, we are now observing a large expansion in the demand for DLs. Research today is often a collaborative effort carried out by groups belonging to different organisations spread worldwide. Motivated by a common goal and funding opportunities, these groups dynamically aggregate into virtual research organisations that share their resources, e.g. knowledge, experimentation results, or instruments, for the duration of their collaboration, creating new and more powerful virtual research environments. These virtual research organisations, set up by individuals that do not necessarily have great economic power or technical expertise, more and more frequently require DLs as tools for accelerating the achievement of their research results. This new potential audience demands less expensive and more dynamic DL development models. They want to be able to set up new DLs that serve their needs for the duration of their collaborations in an acceptable timeframe and with an acceptable cost. The current DL development

model cannot satisfy this large demand; a radical change is needed if we want to be able to address these new emerging requirements.

A great contribution towards the satisfaction of all the above-mentioned requirements can certainly come from the introduction of mechanisms that support a controlled sharing of resources among different organisations. Sharing in this context is not only applied to repositories of content, as is usually meant today, but can be extended to any type of resource needed to build a DL, i.e. language and ontology resources, applications, computers and even staff with the necessary skills for supporting the DL development, deployment and maintenance. Supporting this type of sharing requires the introduction of appropriate solutions at both the technological and organisational levels. These two levels are not independent; instead they strongly influence each other. In fact, the availability of a good technological solution favours the creation of an appropriate organisation, and vice-versa, a successful organisation stimulates the development of new supporting technologies.

In the next section we present the DILIGENT infrastructure as an example of a technological solution for these new DLs. The organisational aspects stimulated by the introduction of this technology are briefly discussed afterwards.

DILIGENT

DIgital Library Infrastructure on Grid ENabled Technology (DILIGENT)[2] is a three-year integrated project (2004-2007) funded by the European Commission under the Sixth Framework Programme for Research and Technological Development. The objective of this project is to develop a digital library infrastructure that will enable members of dynamic virtual research organisations to create on-demand transient digital libraries that exploit shared resources. Resources in this context are multimedia and multi-type content repositories, applications, and computing and storage elements. Following the understanding of DLs expressed in Borgman *et al.* (2002), this project focuses on the development of DLs that "are not ends in themselves; rather they are enabling technologies for digital asset management, electronic commerce, electronic publishing, teaching and learning, and other activities" (p. 7).

From an abstract point-of-view, the DILIGENT infrastructure can be understood as a broker serving DL resource providers and consumers. The providers are the individuals and the organisations that decide to publish their resources under the supervision of the broker, according to certain access and use policies. The consumers are the user communities that want to build their own DLs. The resources managed by this broker are content sources (i.e. repositories of information searchable and accessible through a single "entrance"), services (i.e. software tools that implement a specific functionality and whose descriptions, interfaces and bindings are defined and publicly available) and hosting nodes (i.e. networked entities that offer computing and storage capabilities and supply an environment for hosting content sources and services). Providers register their resources and give a description of them by exploiting appropriate mechanisms provided by the infrastructure. The infrastructure also automatically derives other properties of the resources that are used to enrich the explicit description. The infrastructure manages the registered resources by supporting their discovery, monitoring and usage, and by implementing a number of other functionalities that aim at realising the required controlled sharing and quality of service. A user community can create one or more DLs by specifying a set of requirements. These requirements specify conditions for the information space (e.g. publishing institutions, subject of the content, document types),

for the operations that manipulate the information space (e.g. type of search, tool for data analysis), for the services for supporting the work of the users (e.g. type of personalised dissemination, type of collaboration), for the quality of service (e.g. configuration, availability, response time), and for many other aspects, like the maximum cost, or lifetime. The broker satisfies the community's requirements by selecting, and in many cases also deploying, a number of resources among those accessible to the community, gluing them appropriately and, finally, making the new DL application accessible through a portal. The composition of a DL is dynamic since the DL broker continuously monitors the status of the DL resources and, if necessary, changes them in order to offer the best quality of service. By relying on the shared resources many DLs, serving different communities, can be created and modified on-the-fly, without big investments and changes in the organisations that set them up.

In order to support the transactions between the providers and the consumers, the DILIGENT infrastructure exploits the virtual organisations (VOs) mechanism that has been introduced in the Grid research area (Foster *et al.*, 2001). This mechanism models sets of users and resources aggregated together by highly controlled sharing rules, usually based on an authentication framework. VOs have a limited lifetime, are dynamically created and satisfy specific needs by allocating and providing resources on demand. Through the VOs mechanism the DILIGENT infrastructure glues together the users and the resources of a DL.

Figure 1 graphically illustrates the role of VOs in supporting the brokerage model. The consumers, i.e. the user communities that require DLs to support their needs, are on the left of the figure. The providers, each of which makes a number of resources available, are on the right. The infrastructure acts as a mediator by maintaining a framework where multiple virtual organisations, active on the same shared resources, can co-exist.
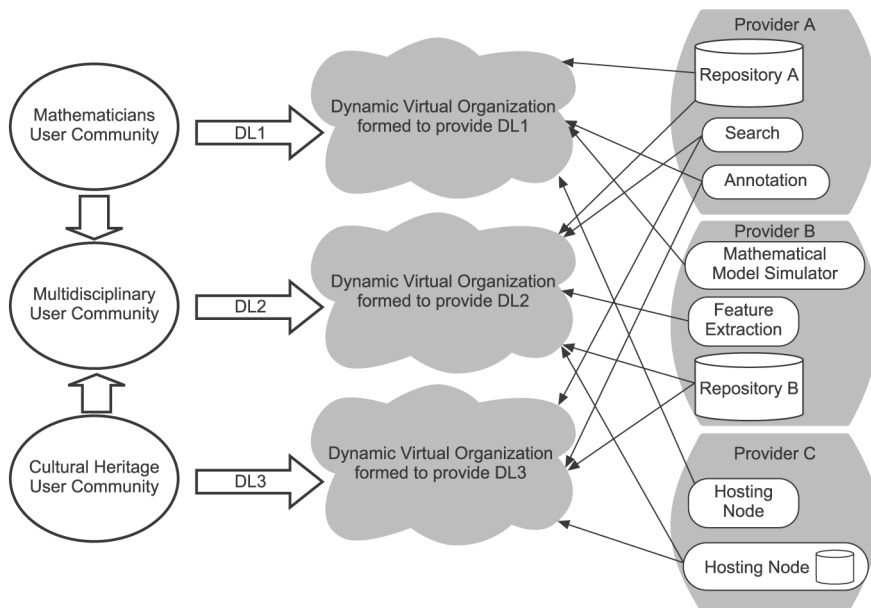

Figure 1.
The role of virtual
organisations in
DILIGENT

The DL development model proposed by DILIGENT is radically new. Within the described framework each DL consumes the required resources only for the time it needs them. This opens a lot of new opportunities for the creation of the functionalities required by the new "knowledge commons" environments. In particular, the exploitation of more effective, but also very computationally expensive algorithms becomes viable at an acceptable cost for many communities. For example, thanks to sharing, the use of the high process-consuming algorithms that automatically extract features from multi-media objects can be exploited in a large number of DLs. Moreover, in the framework established by the new development model, the user communities can easily, and in a timely manner, create and maintain their own DLs with limited resources since the management of the DL is automatically and transparently carried out by the infrastructure.

The system that implements the functionality of the DILIGENT infrastructure is being built by integrating DL and grid technologies (Foster and Kesselman, 2004). The motivation for this design choice relies on the similarity between many of the problems encountered through our new notion of DLs and the issues addressed by the most recent research in the grid domain.

From the functional point-of-view, the DILIGENT system is divided into five functionality clusters:

(1) DL creation and management is responsible for the dynamic construction and maintenance of the transient DLs and for the controlled sharing and management of the resources that are used to implement them. The functionalities offered by this cluster allow users to express the requirements that the DL must fulfil. Moreover, they automatically identify and arrange the pool of resources needed to satisfy these needs.

(2) Content and metadata management implements the handling of DL content and related metadata, the consistent and distributed management of annotations, and the integration of external content and metadata sources.

(3) Process management manages the creation of user processes composed of existing services, the validation of their correctness, the automatic optimisation of their definition according to the resources available and the service characteristics, and their reliable execution. Thanks to this feature, the DILIGENT system can easily be enriched with additional operational workflows to meet new user requirements.

(4) Index and search management is responsible for enabling cost-efficient search and retrieval of information in DLs, while satisfying the level of quality required for the overall data retrieval and delivery operations.

(5) Application specific functionality provides the functionality needed to support user-specific scenarios, like portals, document visualisation, or features extraction.

From the architectural point-of-view, the DILIGENT system is designed as a web services resource framework (WSRF) application (Foster *et al.*, 2004) built on top of the gLite grid middleware[3] released by the Enabling Grids for E-science in Europe (EGEE) project[4]. gLite hides the heterogeneous nature of the computing elements (i.e. services representing a computing resource) on the one hand and storage elements (i.e. services representing a storage resource) on the other hand by providing an environment that facilitates and controls their sharing.

The DILIGENT services are being initially deployed on a project-proprietary gLite infrastructure. Architecturally, this infrastructure is completely interoperable with the EGEE infrastructure. EGEE is currently the largest European grid infrastructure ever built. A number of other recently funded projects will extend this infrastructure to other geographic regions, like Mediterranean countries, Latin America, or China. The interoperability with the EGEE infrastructure will allow any authorised virtual community that wishes to create DLs to also exploit the resources made available by this vast grid infrastructure.

During the project timeframe, the DILIGENT infrastructure will be populated with a number of important archives and software applications provided by the two communities that are participating in the experimentation with the results of the project, one from the environmental e-science domain and one from the cultural heritage domain. The first community is Implementation of Environmental Conventions (ImpECt) and includes leading players in the environmental sector. This community will use DILIGENT to support the organisation of conferences and the preparation of projects and periodical reports. Through DILIGENT this community expects to improve accessibility, interoperability and usability of environmental data, models, tools, algorithms and instruments, integrating the distributed data sources with specialised data handling services. The second community, ARTE, is a community of scholars located in different parts of the world, working together to establish a new discipline that merges experiences from research in medicine, humanities, social sciences and communication. In order to achieve their objectives, these researchers require instruments to ease the construction of multimedia artefacts and to improve support for education.

At the time of the writing of this paper (February 2006), initial experimentation with the features of a DILIGENT DL has already been conducted by implementing simplified services for preparing environmental reports, as required by the ImpECt environmental agencies. Through the exploiting of rich information sources, ranging from raw data sets to maps and graphs archives, these agencies periodically prepare reports on the status of the environment. Currently, this task is performed by first selecting the relevant information from each of the multiple and heterogeneous sources available, then launching complex processing on large amounts of data to obtain "products", like graphs, tables and other summarised information and, finally, producing the required report by assembling all the different parts together. This process, which is repeated periodically, requires a lot of work due to the complexity of interfacing the different sources and tools. Despite the effort expended, the resulting reports do not completely fulfil the requirements of their readers, who would like to have a picture of the environmental status that is updated at the time the report is accessed. A DILIGENT DL offers a more effective framework for the creation and maintenance of these reports. In our experimentation, for example, we have built a DL that exploits content maintained in both repositories including textual documents and archives of Earth observation raw data provided by the European Space Agency. In this DL, which is accessible through a single user interface, all the different kinds of information necessary for creating the reports can be found. By combining this information and by defining how to derive the associated "products" (images, tables, or graphs) from raw data, the users can create their reports much more easily. Moreover, a specialised user interface allows authorised users to access these composite reports by choosing static or dynamic generation. The selection of dynamic generation triggers associated process workflows that, by combining appropriate applications, generate

the required products on demand by processing both the raw data and other intermediate products stored in the DL repositories.

The dynamically generated products are obtained by running those applications that are computationally intensive on the grid. In this way the complex processes required to generate the products are executed in few minutes at a limited cost to the community that is exploiting them. In order to obtain the same performance without the grid, an institution would have to equip its digital library with a great number of computers, while in the case of DILIGENT the institution can also exploit computer capabilities made available by third party organisations. The same is true for storage capacity. Maintaining raw data, intermediate products and high-resolution images requires a large amount of storage capacity. By exploiting the Grid technology, part of this information, especially the temporary part, can be maintained in third-party storage systems.

The role of libraries in future DLs

In the framework envisaged by DILIGENT, libraries play an important role at the organisational level. In particular:

- As providers of resources, they can help to enhance the amount of available resources by making stakeholders aware of the importance of sharing. In particular, as far as the sharing of content is concerned, they can operate by promoting digitisation campaigns and the open access approach. These actions may result in a vast amount of new digital information accessible online which can be exploited by advanced services.

- Within a digital framework, libraries are certainly the best candidates for carrying out content description, maintenance and preservation of resources. By exploiting their large experience acquired in the past, they can contribute to the long-term availability and to the quality of the resources disseminated by the DLs.

- Long-term availability also requires the implementation of models able to support the sustainability of the resources provided. Libraries, either alone or as members of library consortia, can also act as the organisations deputed to define and put in place these models.

- As main resource providers, libraries can work jointly on the definition of common policies and standards. An agreement on these aspects would strongly contribute towards facilitating the design and development of the new complex services required to fulfil the emerging user needs.

- In the future envisaged by DILIGENT, libraries can also play an important role as mediators between the infrastructure and the user communities. In particular, they can proactively promote and facilitate the creation of DLs that respond to the needs of the user communities. They can also assist users by providing, if necessary, the skills required to select, update and exploit the DL content and services.

Concluding remarks

This paper has introduced a vision of the DLs of the future and it has presented DILIGENT, a new technology that is being developed to support this new vision. It has

also discussed the role which libraries can play in the framework envisaged by DILIGENT by outlining, especially, their contribution at the organisational level.

At the time of the writing of this paper, the DILIGENT project is halfway towards the achievement of its objectives. As outlined in the paper, few concrete experiments have already been done to test new functionality and the impact that this new technology may have on supporting the work of specific user communities. The reaction of the user communities involved in these experiments is very encouraging. Other communities, especially scientific communities, have asked to test the technology developed since they found that the DL model proposed can dramatically change the way in which their activities are conducted.

As outlined at the beginning of this paper, the technology by itself is not sufficient to implement the new model envisioned by DILIGENT. A consistent effort is also needed at the organisational level. It is on this level that libraries can play a key role by bringing their valuable experience to bear in the new scenario.

Notes

1. For example the ACM Digital Library: http://portal.acm.org/dl.cfm

2. www.diligentproject.org

3. http://glite.web.cern.ch/glite/

4. http://public.eu-egee.org

References

Borgman, C., Sølvberg, I. and Kovács, L. (Eds) (2002), *Proceedings of the Fourth DELOS Workshop, Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics, Budapest, 6-7 June 2002*, available at: www.sztaki.hu/conferences/deval/presentations/DELOSWorkshop4OnEval_report.pdf (accessed 4 May 2006).

Foster, I. and Kesselman, C. (Eds) (2004), *The Grid: Blueprint for a Future Computing Infrastructure*, 2nd ed., Kaufmann, Amsterdam.

Foster, I., Kesselman, C. and Tuecke, S. (2001), "The anatomy of the grid: enabling scalable virtual organizations", *The International Journal of High Performance Computing Applications*, Vol. 15 No. 3, pp. 200-22.

Foster, I., Frey, J., Graham, S. and Tuecke, S. (Eds) (2004), "Modeling stateful resources with web services", Version 1.1, 3 May, white paper, available at: www-128.ibm.com/developerworks/library/ws-resource/ws-modelingresources.pdf (accessed 4 May 2006).

Ioannidis, Y. (2005), "Digital libraries from the perspective of the DELOS Network of Excellence", in *Proceedings of the IEEE-CS International Symposium: Global Data Interoperability – Challenges and Technologies, 20-24 June, 2005, Sardinia, Italy*, pp. 51-5, available at: http://globalstor.org/ (accessed 4 May 2006).

Leiner, B.M. (1998), "The NCSTRL approach to open architecture for the Confederated Digital Library", *D-Lib Magazine*, Vol. 4 No. 11, available at: www.dlib.org/dlib/december98/leiner/12leiner.html (accessed 4 May 2006).

Corresponding author

Donatella Castelli can be contacted at: donatella-castelli@isti-cnr.it

Strategic planning and customer intelligence in academic libraries

Reinhold Decker and Michael Höppner

Department of Business Administration and Economics, University of Bielefeld, Bielefeld, Germany

Abstract

Purpose – The purpose of this paper is to refer to a current discussion on the effectiveness and efficiency of Bielefeld University Library and concentrates on requirements and conditions of implementing customer intelligence in academic libraries. Moreover, a conceptual framework for a library management information system based on a data warehouse that links external and internal data to support strategic planning processes is introduced.

Design/methodology/approach – Content-related and technical aspects of customer intelligence in academic libraries are outlined and analogies are drawn to commercial enterprises to motivate the conceptual reflections. The paper closes with two examples that demonstrate how multifaceted the data pool for customer intelligence can be in librarianship.

Findings – The paper sensitizes to the advantages of systematically generating customer knowledge in academic libraries for strategic planning and customer orientation.

Practical implications – The suggested approach can serve as a basis for the development of data-based decision support systems focusing on the tracking of the usage of library services and customer preferences over time.

Originality/value – Up to now the discussion of customer intelligence as a foundation of strategic planning in academic libraries has been almost a blank space in the literature. The paper contributes to fill this gap.

Keywords Intelligence, Library management, Strategic planning, Academic libraries

Paper type Conceptual paper

Introduction

Though it is almost a truism that libraries have to rethink their services and their position in the internet age, there are some obvious reasons to concentrate on strategic planning issues based on customer intelligence in the following article. Since nowadays information seems to be at everyone's fingertips very easily, the role of libraries in the information chain is no longer accepted without question. Even the patrons of academic libraries do not generally realize that scientific e-journals available through the university's intranet are not for free. Only a few customers are really aware of the financial efforts and know-how which are necessary to build up state-of-the-art information systems that seamlessly integrate bibliographic records, electronic full texts, and document delivery services, all of which stem from various sources. This diminishing awareness of libraries could even be seen as a positive aspect of an adequate but not obtrusive service. However, it becomes crucial with respect to actual budgetary constraints as well as with respect to modernistic but not always appropriate attempts to reorganize the information, communication, and media services of a university (Hanson, 2005).

Today, budgetary decisions are based on an input-output orientation. So, libraries have to supply evidence that their contribution to the output of the university as a whole, i.e. to progress in teaching and research, at least matches their consumption of resources. This, of course, is difficult, because the input is measured in Euros, but the output is not. Moreover, since the budgetary decisions are made by the customers, it has become evident that all strategic considerations of libraries have to aim at customer satisfaction and at the economical consumption of resources. Therefore, the effectiveness and efficiency of services are becoming a ubiquitous challenge in library service planning. The crucial question is whether advanced, but cost-intensive state-of-the-art services should be provided in the future or whether traditional, but appropriate services have their place as well? Thus customer intelligence (Kelly, 2006) becomes an essential part of strategic planning. This recognition is not really new. In fact, "user research" has a long tradition in German libraries. But up to now, this was primarily done by public surveys focusing on existing services. Nowadays, there are also some more advanced approaches like the gap analysis of services, as supported by LibQUAL + (Cook, 2002), or the prospective preference measurement by means of conjoint analysis (see below).

Strategic planning at Bielefeld University Library

Most customers have a rather traditional understanding of libraries and their services. Bielefeld University Library, for instance, is ranked high by its customers in general, but this often primarily results from its outstanding service hours, instead of being based on what is essential in our own understanding, particularly our leading role in developing and establishing advanced electronic library services. As an example, our academic search engine Bielefeld Academic Search Engine (BASE) (Summann and Lossau, 2004), which is ranked highly by library experts, is graded rather low by our customers in comparison to more traditional retrieval instruments like the Online Public Access Catalogue (OPAC). This traditional view conflicts with the obvious integration of internet-based search engines like Google in every-day life. One reason for this contradiction might be insufficient information literacy on the part of the customers, who often cannot totally assess the completeness, the relevance, and the reliability of information "googled" from the internet. Therefore, attitudes and opinions of customers regarding library services cannot be the only foundation for new service development, but have to be accompanied by alternative expertise and by convincing marketing activities.

Nowadays, management techniques like mission statements, catalogues of services, and service level agreements are already well-established in libraries for strengthening the overall service orientation. This is the way Bielefeld University Library has been following since 2004 for discussing its effectiveness. For pragmatic reasons, we are doing without an explicit mission statement for the moment, but we refer to an idea by Schelsky (1967), who stated as an essential for Bielefeld University and its library that the value of a library is not measured by the bulk of books on the shelves, but by the speed the customer can be supplied with the books he is really needing (p. 65). We are strongly convinced that this formerly provocative idea can still be adequately interpreted in the Internet age. Therefore, it is taken as the mission idea underlying the catalogue of services developed for Bielefeld University Library.

Service orientation cannot really be strengthened by management directive, but rather takes effect only through the involvement of all employees. For this reason, Bielefeld University Library started the discussion by screening the library environment for new service ideas, e.g. using expertise from our employees, from external experts, and from the literature. A list of about 250 services ranked essential for the future constituted the basis of a comprehensive conjoint analysis study to measure the preferences of our customers (Decker and Hermelbracht, 2006). In addition, four internal working groups developed first drafts of detailed product sheets for library services, which should establish a catalogue of services later on. The whole process was based on detailed information and fruitful discussions during several staff meetings and it was coordinated by a steering group. The involvement of the staff council from the very beginning proved to be very helpful. Moreover, the outstanding engagement of a great number of our employees in these discussions was very encouraging. The outcome from the working groups was a set of about 200 drafts building a solid basis for strategic planning and the optimization of internal transactions. Then, the drafts were rearranged by an editorial group to some 50 product sheets, which were rated to be of obvious importance for customer orientation. Finally, both the optional and the existing services were ranked once more by the library employees to validate the product ratings. This additional evaluation step was confirming the picture to a large extent, and we will be able to publish the product sheets soon.

So far only the effectiveness of our services has been considered. Therefore, the next step will deal with their efficiency. Strengthening the efficiency should not be confused with pure savings of costs or even of staff. Of course, a "one-person-library" might be very cost-efficient, but its service efficiency would exponentially decrease with the number of service requests. So, cost minimization alone cannot be a strategic model for successful academic libraries. In fact, beyond all possible options for rationalization there is an indissoluble connection between the quality of services and the resources needed. Thus, the decision of how to adjust the level of services against the level of costs is not only economical but rather political.

Efficiency may be strengthened by organizational development and development of personnel, as well as by reengineering business and by readjusting objectives. Such changes may not have been intended as independent aims from the very beginning but they can arise from common discussions. In contrast to discussing effectiveness, where the whole spectrum of services has to be considered, we will only be able to discuss the topics that are most important for service efficiency. Firstly, because of the traditional understanding of libraries by our customers, we have to consider all our customer-supporting activities, including marketing. Secondly, we have to evaluate critically our workflow and our transactions with respect to established library performance indicators such as the German library index BIX, which points out some already known but not completely understood deficits. Last but not least, we have to check our data processing routines, which are an indispensable foundation for most of our services, but have meanwhile gained a complexity that seems to be no longer of manageable size and that should be reduced if possible. Data processing has also been considered as a foundation of strategic planning where sophisticated data analysis techniques are needed that enable problem-oriented decision support in academic library management.

Customer intelligence and decision support

To know the customer is – as already indicated in the previous sections – a prerequisite of and the key to customer orientation. So far it is not really surprising that customer intelligence is not only intensively discussed in commercial enterprises, but also arouses increasing interest in academic libraries. But what does customer intelligence (CI) actually mean? In business CI primarily aims at gaining a comprehensive understanding of customers and their behavior by means of intelligent tools, which enable a more pointed customer contact and a higher degree of customer loyalty. Though this understanding of CI explicitly emphasizes the relevance of method-based approaches, the general relevance of intuition and pure experience should not be questioned. On the other hand, the rapidly increasing amount of data in libraries causes the management to apply appropriate techniques for information production. Typical questions to be answered in CI processes are:

- Who are the customers and what needs and preferences do they have?
- How and when do they use the library services?
- How loyal are the customers and who are the "valuable" ones?
- What activities or offerings are most effective in generating customer satisfaction?

The topic CI is frequently mentioned in the same breath as customer relation management (CRM). The underlying assumption is that an enterprise can only be successful in the long term if it is managed in a market or rather customer-oriented way. As a consequence, the success of any particular CRM activity is largely dependent on the ability to organize and deploy the usually diverse information sources of the library towards a common purpose: managing the customer relationship in all of its forms. Therefore, effective CRM is not possible without detailed and up-to-date customer knowledge.

Unfortunately, there is neither a unanimous understanding of CI in marketing science nor a well-defined set of CI methods. In particular, a validated theory, on the basis of which the substantial elements of CI could be concretized, is still missing. However, referring to the relevant literature, a certain consensus regarding the instruments needed for implementing CI seems to exist. The availability of a data warehouse as well as a method pool that includes both on-line analytical process (OLAP) and data mining techniques is frequently required. This particularly applies when the data pool to be analyzed for building customer knowledge is large. Thus, the success of data mining processes strongly depends on the data input (Delmater and Hancock, 2001). Library data can be mined if it is in the right format and well integrated.

The basic structure of a library management information system (LiMIS), as it may be implemented in the CI context, is depicted in Figure 1. The heart of this LiMIS is a data warehouse. The basic idea behind this is to transfer the data generated by different operational systems to an autonomous database, the data warehouse. The term "operational system" summarizes all kinds of data sources within a library, e.g. the loan terminals, the library website and even periodically conducted user surveys. By separating the original data sources from the LiMIS, the justified demand for anonymity on the part of the library users can be taken into account. In this sense the
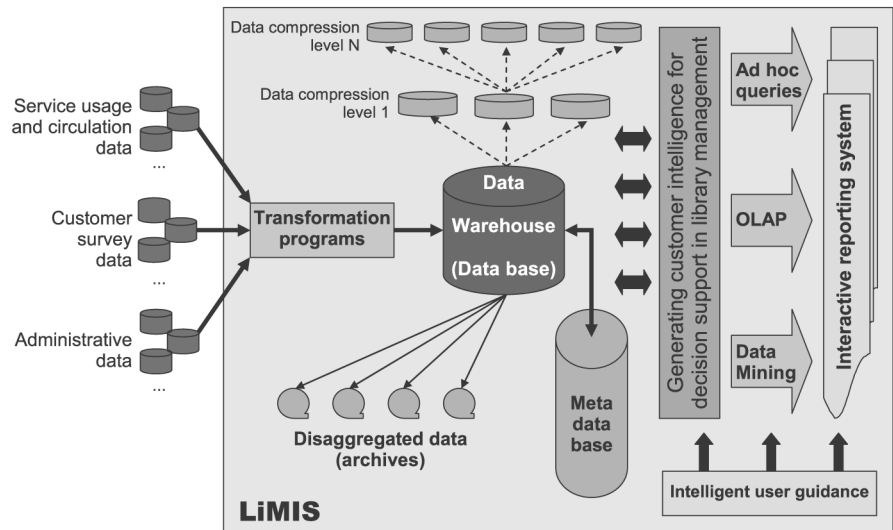
transformation interface not only serves as a data converter but is also used to make the data anonymous before they are made accessible for the LiMIS user. By this means it can be ensured that those persons who consult the system for decision support purposes do not have access to any data that would allow the identification of the individual library user. In fact, the focus is on the provision of individualized but not personalized customer knowledge. In the best case, the use of the analytical instruments is managed by an intelligent front-end, as is already becoming a standard in statistical analysis systems such as SAS.

The basic relevance of data warehouses in practice was recently supported by an empirical study carried out by Ariyachandra and Watson (2006). The focus of this study was on the general success of data warehousing and business intelligence activities in companies. The 454 companies included in this survey mostly rated their implementations as tending towards success with regard to the attainable decision support. But this study also provides evidence that the introduction of a data warehouse needs considerable resources and cannot be done "by the way". The comparatively high total costs diagnosed for companies also apply to libraries, at least as long as adequate standard software packages for holistic implementations do not exist.

Building a data warehouse for generating customer intelligence inevitably leads to the question of how to obtain relevant customer information. In order to answer this question, a closer look at the customer process itself may be helpful. Usually four phases are distinguished. In the search and orientation phase (1), we can learn something about the things the customers are interested in. Information of this kind can be gained by analyzing the individual usage of a library's website. The obvious starting point for this is the log file data which is automatically generated when a user visits and surfs a website. The "purchase" or rather library usage behavior (phase 2) finds its expression, for example, in individual lending data or an observable stay in the library. The latter would be possible by using electronic visit counters at the library

entrances, for example. The development of satisfaction and loyalty in phase 3 is reflected in regular transaction data and appropriate customer surveys. Finally, if satisfaction is high, the repeated usage of the respective library services (phase 4) can be measured according to phase 3. With these aspects in mind we can have a closer look at the CI process in libraries, which can be divided into six steps:

(1) Capturing customer data across all points of interaction between the customers and the library according to the decision preferences of the library management.

(2) Integration of the anonymized customer data into the data warehouse to enable trans-sectoral analyses at low IT costs and according to relevant data protection law.

(3) Application of advanced data analytical techniques to generate customer information with predictive potentials.

(4) Transformation of the customer knowledge into customer-oriented library services.

(5) Measurement of the incremental benefits resulting from particular customer investments.

(6) Adjustment and/or refinement of the CI process to improve future efforts.

So, data warehouse and OLAP/data mining techniques build the "bridge" between customer data and customer information, while customer knowledge is the result of the translation and dissemination of customer information in the library. Customer knowledge can be explicit (in the form of structured information in the LiMIS) or implicit (in the form of knowledge in the mind of library employees or the customers themselves) (Rollins and Halinen, 2005). In practice, both forms of customer knowledge have to complement each other if the aim is maximum impact of the related (marketing) activities. In what follows, we are going to sketch significant challenges in developing a LiMIS intended to be used to support the process of generating customer intelligence in libraries, whereby the focus will be on the explicit form of customer knowledge.

Selected issues of developing a LiMIS

The development of a LiMIS for CI activities and decision support in strategic library planning involves the solving of a couple of more or less complex subtasks. First, the relevant data structures have to be identified. That is to say, we have to answer the question of which data (generated by the different operational systems being employed within and in relation to the library) should be represented in the data warehouse. This, on the other hand, leads to the crucial question of how to deal with the heterogeneity of different library data, starting with quantitative loan data and ending up with qualitative user survey data (e.g. attitudes and opinions with respect to new library services). Therefore, the successful implementation and application of a LiMIS largely depends on the adequacy of the underlying data model. To ease the data transformation process, a flexible data model is needed which allows the description of the relevant data in a consistent and largely standardized way.

One possibility is to use an ASCII-based data interface (Baier and Marx, 1992). The idea is to describe all data considered to be relevant with an a priori defined set of

attributes. The following example demonstrates the procedure by referring to an OPAC search protocol (Table I).

A seven-attribute data set to be transmitted weekly to the data warehouse, for instance, might then look like this:

7 text *login* date *date of OPAC search* time *time of OPAC search* text *search mode used* integer *number of hits* text *author of the document* text *title key word*
"internal" 2006.01.04 15:13:11 "1" 1 # "guatemala politicos"
"external" 2006.01.04 15:19:04 "2" 6 "Murray" "Econometrics"
"internal" 2006.01.04 15:22:02 "2" 2 "Burns & Bush" "Marketing Research"
"external" 2006.01.04 15:23:12 "2" 0 # "Goethe-Institut"
"internal" 2006.01.04 15:24:19 "1" 1 "Hardgrove" "Mathematics Library"

The attribute *login* informs about the access to the OPAC system, whereas the nominal values "1" and "2" for attribute *search mode used* distinguish between single-line search and field search. With the attributes *author of the document* and *title key word* the content of the search has been specified. The remaining attributes are self-explanatory, and the symbol # marks a missing value. An ASCII-based data interface does not only ease the integration of different operational systems, but also enables extensions of the data model with acceptable efforts. In this context, the question of the adequate level of automation regarding the data transformation process arises. A high level of automation reduces the daily operating costs, but also requires anticipatory action when developing the data model in order to avoid the risk of costly subsequent improvements or corrections.

A further crucial point concerns the actual information that should be provided by the LiMIS. Determining practical system requirements virtually means referring to two aspects, namely the type of information to be provided and the level of aggregation desired. The first aspect can be taken into account by involving the library management in both the design and the implementation process. Kick-off meetings before, and evaluation workshops during the development process, focusing on the

Number of attributes used for describing the data set,	e.g. 7
Identification of attribute type 1,	e.g. text
Description of attribute 1,	e.g. *login*
Identification of attribute type 2,	e.g. date
Description of attribute 2,	e.g. *date of OPAC search*
Identification of attribute type 3,	e.g. time
Description of attribute 3,	e.g. *time of OPAC search*
Identification of attribute type 4,	e.g. text
Description of attribute 4,	e.g. *search mode used*
Identification of attribute type 5,	e.g. integer
Description of attribute 5,	e.g. *number of hits*
Identification of attribute type 6,	e.g. text
Description of attribute 6,	e.g. *author of the document*
Identification of attribute type 7,	e.g. text
Description of attribute 7,	e.g. *title key word*
value_attribute_1 value_attribute_2 value_attribute_3	
value_attribute_4 value_attribute_5 value_attribute_6	
value_attribute_7	
value_attribute_1 value_attribute_2 value_attribute_3 etc.	

Table I.
Set of attributes for describing an OPAC search protocol

information needs the LiMIS should be able to meet and the level of detail required for the respective decision topics, can significantly increase acceptance. Typical questions to be answered are:

- What kind of information plays a major role in decision making in library management?
- Which decisions would particularly benefit from the availability of detailed customer information?
- In what order is the relevant information entering the decision processes?
- Which management figures should be made available by the LiMIS reporting system?

But sound planning and a deliberate balance between technically possible and actually useful options are also indispensable regarding the method selection and implementation process. The method selection problem primarily concerns the data-mining component (see Figure 1). To avoid needless implementations of partly sophisticated techniques, a goal-driven procedure seems to be favorable. That is to say the decision or planning process to be supported by the LiMIS determines the (data mining) method to be integrated. If the library management is interested in analyzing the conjoint demand of documents (e.g. in the form of textbooks), for example, the application of an association rules-based approach (Agrawal and Srikant, 1994) to loan or OPAC search data suggests itself. Customer knowledge of this kind can be helpful in acquisition planning, for instance. Budget allocation analyses to objectify future library budgets on the departmental level, however, may be carried out by means of logistic discriminant analysis based on the intensity of media usage measured in the respective departments. Already, these two examples are showing that the method normally has to follow the decision problem, and not the reverse. By first determining the decision context to be covered with the LiMIS, and then integrating those methods that seem to be most suited for solving the related data analysis problem, a discouraging method overload can be avoided. Here, the expert system discussion of the late 1980s and early 1990s of the last century comes to mind again. At that time the euphoria in science quickly gave way to a certain disillusionment, when the developers had to recognize that the automation of decision processes is neither trivial nor necessarily accepted on the part of the users. To prevent similar effects, data mining should be done in a goal-driven manner, especially in fields of application, where the regular use of this type of data analysis is not yet widespread.

For similar reasons, the design and implementation of the reporting system also deserves closer attention. In the ideal case, the relevant customer information (the figures and tables as well as the related interpretations) is provided at the touch of a button. To ensure that the reporting system always delivers its results at the optimal level of aggregation and detail, two aspects are most notable, namely the basic importance of a particular aspect for the decision problem at hand, and the way the interestingness of an available result is determined. The first point means that, depending on the decision problem and the customer knowledge available already, the LiMIS users might be interested in different levels of information aggregation. Strategic decisions in the budgeting context mentioned above, for example, would require a more global view, e.g. with a general focus on the quarterly budget available

or needed for the electronic journals subscribed by the individual departments of a university. On the other hand, a concrete budget discussion with the head of a particular departmental library, with respect to the extension or cancellation of a journal subscription, would require detailed information about previous utilizations of the respective journal by the department members. The selection of the adequate level of aggregation can be supported by an intelligent (e.g. rule-based) user guidance system, which determines the level by means of a goal-driven dialogue to specify the individual concern. The second point refers to a more technical problem in data mining. Although data mining is often defined as the – more or less – automatic search and identification of patterns in large amounts of data, this process is still controlled by the user and the results are not inevitably self-explanatory. If association rules are used to uncover interesting patterns in individual textbook-loaning behavior, for example, the results (i.e. the detected rules) are rated using so-called measures of interestingness, such as the lift and the conviction (Brin *et al.*, 1997). Due to their different definitions, these measures may lead to different assessments. To be able to evaluate the virtual relevance or interestingness of a pattern, users who are not familiar with the respective methodology need specific support regarding the interpretation of the individual measures. Here, once again, intelligent user guidance is indispensable, if the data mining process is not to degenerate into a randomized picking of data.

Data and methods for generating customer intelligence
The following two examples show how different the data and methods that may define the basis of CI can be in practice. We start with so-called preference data generated by means of conjoint analysis. The basic idea of conjoint analysis is to decompose rank-ordered evaluation judgments of a set of alternatives (e.g. products) into relevant components based on qualitative attributes. That is to say, each alternative is considered as a combination of attributes, each of them contributing to its utility to some degree. For each level of each attribute, a numerical part-worth utility value is computed from the rank orderings that have been collected from the respondents in an appropriate survey. The sum of the part-worth utility values for an alternative is assumed to be an estimate of its unknown (latent) utility. The goal is to compute the part-worth utility values such that the computed utilities of the considered alternatives reflect the original rank orderings of the alternatives considered as accurate as possible (see Decker and Hermelbracht, 2006, for methodical details). Figure 2 illustrates the preferences, or rather part-worth utility values, regarding future options for lending and delivering books at Bielefeld University Library. The study underlying this example comprises the answers of about 2,100 library users. From Figure 2 we learn, for example, that the current practice at Bielefeld University Library is preferred most, whereas a combination of conventional loan (CL) with a chargeable home delivery service would obtain comparatively low approval.

The data underlying Figure 2 have been generated especially for preference measurement purposes and thus represent typical primary data. In contrast to this, the second example is based on secondary data (i.e. on data that have been collected for some purpose other than the one considered now). It demonstrates the application of a pattern mining technique to media usage data. In the respective study about 3,770 book profiles were analyzed using a self-organizing neural network approach. The profiles had been defined by means of 34 items describing the usage of the respective books in
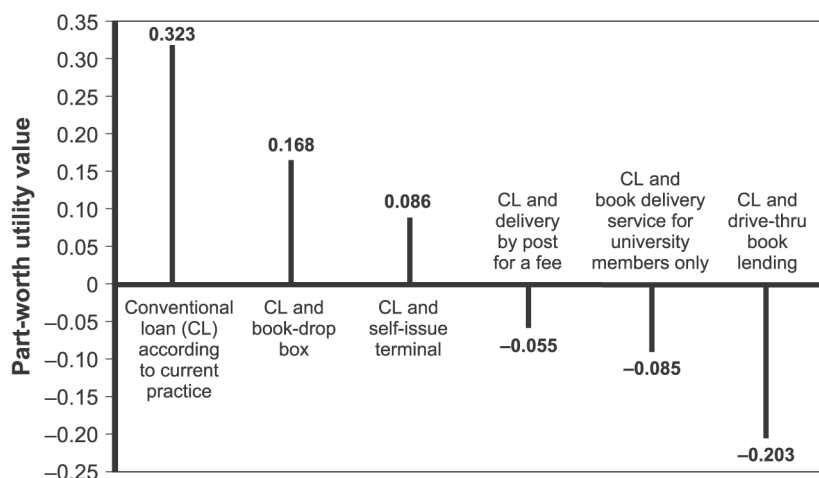
Figure 2.
Part-worth utility values
of alternative options
related to book loan and
delivery

the considered period of time. As a result, 11 book usage patterns could be identified. Table II displays profile information for eight items of two book types (the numerical values equal the weights defining the neurons of the neural network). Books of type 2 – on average – feature about 375 pages and are mostly available in the second or even a higher edition. The total number of loans per term is higher than five (5.25) and the average number of biweekly loans equals 5.19. Looking closer at the authors of this type of book, and/or the topics they are focusing on, may be helpful in future acquisitions decisions. Books of type 2 may also be candidates for reading lists of related lectures.

Concluding remarks

Finding the truth about customers is an ambitious task, not only in commercial enterprises but also in academic libraries. Management information systems based on data warehouses representing data that can be assumed to explain, or at least motivate, customer preferences and behavior are promising tools for decision support in strategic planning. The crucial point is to find the optimal balance between the required level of detail and justifiable abstraction. Therefore, acceptance of a LiMIS strongly depends on the extent to which its developers succeed in integrating the available methods into concrete decision tasks, or, in other words, the extent to which it helps to associate

Item	Book type 1	Book type 2
Number of available copies	1.20	2.10
Number of pages	250	375
Last edition	1.12	2.25
Year of publication	1991	1994
Total number of loans	0.89	5.25
Number of reservations	0.04	0.66
Number of biweekly loans	0.13	5.19
Number of copies lent during the term	0.57	3.70

Table II.
Patterns in book
usage data

tessellated customer information in order to obtain coherent knowledge about the customer. Thus, meeting the customer intelligence challenge in academic libraries is much more than investing in modern IT technology.

References

Agrawal, R. and Srikant, R. (1994), "Fast algorithms for mining association rules", in Bocca, J. (Ed.), *Proceedings of the 20th International Conference on Very Large Databases, September 12-15, Santiago de Chile*, Morgan Kaufmann, Hove, pp. 487-99.

Ariyachandra, T. and Watson, H. (2006), "Benchmarks for BI and data warehousing success", *DM Review Magazine*, January, available at: www.dmreview.com/article_sub. cfm?articleId = 1044330 (accessed 8 May 2006).

Baier, D. and Marx, S. (1992), "Data management in a knowledge-based system for marketing research", in Schader, M. (Ed.), *Analyzing and Modeling Data and Knowledge*, Springer, Berlin, pp. 189-97.

Brin, S., Motwani, R., Ullman, J.D. and Tsur, S. (1997), "Dynamic itemset counting and implication rules for market basket data", in Peckman, J.M. (Ed.), *proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, May 13-15, Tuscon, ACM Press, New York, NY, pp. 255-64.

Cook, C. (2002), "The maturation of assessment in academic libraries: the role of LibQUAL+™", *Performance Measurement and Metrics*, Vol. 3 No. 2, pp. 34-112.

Decker, R. and Hermelbracht, A. (2006), "Planning and evaluation of new academic library services by means of web-based conjoint analysis", *Journal of Academic Librarianship*, in press.

Delmater, R. and Hancock, M. (2001), *Data Mining Explained: A Manager's Guide to Customer-Centric Business Intelligence*, Butterworth-Heinemann, Woburn, MA.

Hanson, T. (2005), "Twenty years of convergence in the UK", in Hanson, T. (Ed.), *Managing Academic Support Services in Universities*, Facet Publishing, London, pp. 1-9.

Kelly, S. (2006), *Customer Intelligence: From Data to Dialogue*, Wiley, Chichester.

Rollins, M. and Halinen, A. (2005), "Customer knowledge management competence: towards a theoretical framework", in Spraque, R.H. (Ed.), *IEEE – Proceedings of the 38th Annual Hawaii International Conference on System Sciences, January 3-6, 2005, Hawaii*, IEEE Computer Society, Los Alamitos, CA, pp. 240-50.

Schelsky, H. (1967), "Grundzüge einer neuen Universität. Eine Denkschrift", in Mikat, P. and Schelsky, H. (Eds), *Grundzüge einer neuen Universität. Zur Planung einer Hochschulgründung in Ostwestfalen*, Bertelsmann, Gütersloh, pp. 35-70.

Summann, F. and Lossau, N. (2004), "Search engine technology and digital libraries: moving from theory to practice", *D-Lib Magazine*, Vol. 10 No. 9, available at: www.dlib.org/dlib/september04/lossau/09lossau.html (accessed 8 May 2006).

Corresponding author
Reinhold Decker can be contacted at: rdecker@wiwi.uni-bielefeld.de

e-Science and its implications for the library community

Tony Hey

Microsoft Corporation, Redmond, Washington, USA, and

Jessie Hey

*School of Electronics and Computer Science and
University of Southampton Libraries, University of Southampton,
Southampton, UK*

Abstract

Purpose – The purpose of this article is to explain the nature of the "e-Science' revolution in twenty-first century scientific research and its consequences for the library community.

Design/methodology/approach – The concepts of e-Science are illustrated by a discussion of the CombeChem, eBank and SmartTea projects. The issue of open access is then discussed with reference to arXiv, PubMed Central and EPrints. The challenges these trends present to the library community are discussed in the context of the TARDis project and the University of Southampton Research Repository.

Findings – Increasingly academics will need to collaborate in multidisciplinary teams distributed across several sites in order to address the next generation of scientific problems. In addition, new high-throughput devices, high-resolution surveys and sensor networks will result in an increase in scientific data collected by several orders of magnitude. To analyze, federate and mine this data will require collaboration between scientists and computer scientists; to organize, curate and preserve this data will require collaboration between scientists and librarians. A vital part of the developing research infrastructure will be digital repositories containing both publications and data.

Originality/value – The paper provides a synthesis of e-Science concepts, the question of open access to the results of scientific research, and a changing attitude towards academic publishing and communication. The paper offers a new perspective on coming demands on the library and is of special interest to librarians with strategic tasks.

Keywords Digital libraries, Digital storage

Paper type Research paper

Introduction

As Thomas Friedman (2005) eloquently explains in his book *The World is Flat*, the convergence of communication and computing technologies is changing the world of both business and leisure. It would be naïve to think that the academic research community will be immune from these changes. The methodology of research in many fields is changing and we are on the threshold of a new era of data-driven science. In the last few decades computational science has emerged as a new methodology for scientific research on an equal footing with the traditional experimental and theoretical methodologies. Simulation is now used as a standard weapon in the armory of the scientist to explore domains otherwise inaccessible to the traditional research

methodologies – such as the evolution of the early universe, the design of new materials, the exploration of climatology over geological timescales and, of course, the weather forecasts we now take for granted. Its use in industry is becoming even more widespread with computational fluid dynamics and finite element simulations now an essential part of the design process, complementing traditional experimental wind tunnel and safety testing in the aero and auto manufacturing industries, with simulations of oil fields and analysis of seismic data now playing a key role in the oil and gas industry, and with simulation playing an increasingly important role in the drug design life cycle in the pharmaceutical industry.

The next decade will see the emergence of a new, fourth research methodology, namely "e-Science" or networked, data-driven science. Many areas of science are about to be transformed by the availability of vast amounts of new scientific data that can potentially provide insights at a level of detail never before envisaged. However, this new data dominant era brings new challenges for the scientists and they will need the skills and technologies both of computer scientists and of the library community to manage, search and curate these new data resources. Libraries will not be immune from change in this new world of research. The advent of the web is changing the face of scholarly publishing and the role of publishers and libraries. The National Science Foundation Blue Ribbon Report on Cyberinfrastructure lays out a vision of this new world. On publishing, the report states:

> The primary access to the latest findings in a growing number of fields is through the Web, then through classic preprints and conferences, and lastly through refereed archival papers (Atkins *et al.*, 2003, p. 9).

And on scientific data the report states:

> Archives containing hundreds or thousands of terabytes of data will be affordable and necessary for archiving scientific and engineering information. (Atkins *et al.*, 2003, p. 11).

This paper explores some of the challenges facing both the scientific and library communities in this new emerging world of research and delineates the key role that can be played by computer science and by IT companies such as Microsoft in assisting the research community.

e-Science and Licklider's vision

It is no coincidence that it was at CERN, the particle physics accelerator laboratory in Geneva, that Tim Berners-Lee invented the world wide web. Given the distributed nature of the multi-institute collaborations required for modern particle physics experiments, the particle physics community urgently needed a tool for exchanging information. It was their community who first enthusiastically embraced the web as a mechanism for information exchange within their experimental collaborations and it was no accident that the first web site in the USA was at the Stanford Linear Accelerator Center Library. As we all now know, since its beginnings in the early 1990s, the web has not only taken the entire scientific world by storm but also the worlds of business and leisure. Now, just a decade or so later, scientists need to develop capabilities for collaboration that go far beyond those of the original world wide web. In addition to being able just to access information from different sites, scientists now want to be able to use remote computing resources, to integrate, federate and analyze

information from many disparate and distributed data resources, and to access and control remote experimental equipment. The ability to access, move, manipulate and mine data is the central requirement of these new collaborative science applications – whether the data is held in flat files or databases, or is data generated by accelerator or telescopes, or data gathered in real-time from potentially mobile sensor networks.

In the UK, at the end of the 1990s, John Taylor became Director General of Research Councils at the Office of Science and Technology (OST) in the UK – roughly equivalent to Director of the National Science Foundation (NSF) in the USA. Taylor had been Director of Hewlett-Packard (HP) Laboratories in Europe and HP's vision for the future of computing has long been that IT resources will become a new "utility". Rather than purchase IT infrastructure, users will pay for IT services as they consume them, in the same way as the conventional utilities such as electricity, gas and water – and now mobile telephones. In his role at the OST as overseeing the funding of UK scientific research, Taylor realized that many areas of science could benefit from a common IT infrastructure to support multi-disciplinary and distributed collaborations. He articulated a vision for this type of distributed, collaborative science and introduced the term "e-Science":

> e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it (Taylor, 2001).

It is important to emphasize that e-Science is not a new scientific discipline in its own right: e-Science is shorthand for the set of tools and technologies required to support collaborative, networked science. The entire e-Science infrastructure is intended to empower scientists to do their research in faster, better and different ways.

Of course, these problems are not new – the computer science community has been grappling with the challenges of distributed computing for decades. Indeed, such an e-Science infrastructure was very close to the vision that J.C.R. Licklider ("Lick") took with him to Advanced Research Projects Agency (ARPA) when he initiated the core set of research projects that led to the creation of the ARPANET. Larry Roberts, one of his successors at ARPA and principal architect of the ARPANET, described this vision as follows:

> Lick had this concept of the intergalactic network which he believed was everybody could use computers anywhere and get at data anywhere in the world. He didn't envision the number of computers we have today by any means, but he had the same concept – all of the stuff linked together throughout the world, that you can use a remote computer, get data from a remote computer, or use lots of computers in your job. The vision was really Lick's originally. (Segaller, 1998, p. 40).

The ARPANET of course led to the present day internet – but the killer applications have so far been email and the web rather than the distributed computing vision originally described by Licklider. Of course, in the early 1960s, Licklider was only envisaging connecting a small number of rather scarce and expensive computers, and at relatively few sites. However, over the past 30 years, Moore's Law – Gordon Moore's prediction that the number of transistors on a chip would double about every 18 months so that the price-performance is halved at the same time – has led to an explosion in the number of supercomputers, mainframes, workstations, personal computers and PDAs that are now connected to the Internet. Already we are beginning

to see programmable sensors and RFIDs – intelligent tagging devices – being connected to the network.

An example of e-Science: The CombeChem, eBank and SmartTea projects
The CombeChem project[1] was funded by the Engineering and Physical Sciences Research Council in the UK and its goals were to enhance the correlation and prediction of chemical structures and properties by using technologies for automation, semantics and grid computing (see Frey *et al.*, 2003; Hughes *et al.*, 2004). A key driver for the project was the fact that large volumes of new chemical data are being created by new high throughput technologies. One example uses the technologies of combinatorial chemistry in which large numbers of new chemical compounds are synthesized simultaneously. The volume of data and the speed by which it can be produced highlights the need for assistance in organizing, annotating and searching this data. The CombeChem team consisted of a collection of scientists from several disciplines – chemistry, computer science and mathematics – who developed a prototype test-bed that integrated chemical structure-property data resources with a "grid" style distributed computing environment. The project explored automated procedures for finding similarities in solid-state crystal structures across families of compounds and evaluated new statistical design concepts in order to improve the efficiency of combinatorial experiments in the search for new enzymes and pharmaceutical salts for improved drug delivery.

The CombeChem project also explored some other important e-Science themes. One theme concerned the use of a remote X-ray crystallography service for determining the structure of new compounds. This service can be combined in workflows with services for computer simulations on clusters or searches through existing chemical databases. Another important e-Science theme was the exploration of new forms of electronic publication – both of the data and research papers. This e-Publication theme was examined in the eBank project[2] funded by the Joint Information Systems Committee (JISC). One of the key concepts of the CombeChem project was that of "Publication@Source" which establishes a complete end-to-end connection between the results obtained at the laboratory bench and the final published analyses (Frey *et al.*, 2002). This theme is linked to yet another of the e-Science themes explored in the CombeChem project that was concerned with human-computer interfaces and the digital capture of information. In the associated SmartTea project[3], computer scientists studied the way chemists within the laboratory used their lab notebooks and developed acceptable interfaces to handheld tablet technology (see Schraefel *et al.*, 2004a; Schraefel *et al.*, 2004b). This is important since it facilitates information capture at the very earliest stage of the experiment. Using tablet PCs, the SmartTea system has been successfully trialed in a synthetic organic chemistry laboratory and linked to a flexible back-end storage system. A key usability finding was, not surprisingly, that users needed to feel in control of the technology and that a successful interface must be adapted to their preferred way of working. This necessitated a high degree of flexibility in the design of the lab book user interface. The computer scientists on the team also investigated the representation and storage of human-scale experiment metadata and introduced an ontology to describe the record of an experiment.

A novel storage system for the data from the electronic lab book was also developed in the project. In the same way that the interfaces needed to be flexible to cope with

whatever chemists wished to record, the back end solutions also needed to be similarly flexible to store any metadata that might be created. This electronic lab book data feeds directly into the scientific data processing. All usage of the data through the chain of processing is now effectively an annotation upon it, and the data provenance is explicit. The creation of original data is accompanied by information about the experimental conditions in which it is created. There then follows a chain of processing such as aggregation of experimental data, selection of a particular data subset, statistical analysis and modeling and simulation. The handling of this information may include explicit annotation of a diagram or editing of a digital image. All of this generates secondary data, accompanied by the information that describes the process that produced it. This digital record is therefore enriched and interlinked by a variety of annotations such as relevant sensor data, usage records or explicit interactions. By making these annotations machine processable, they can be used both for their anticipated purpose and for subsequent unanticipated reuse. In the CombeChem project this was achieved by deployment of web services and semantic web technologies (Berners-Lee *et al.*, 2001). Resource description framework (RDF) was used throughout the system: at present there are over 70 million RDF triples in the CombeChem triplestore. This system was found to give a much higher degree of flexibility to the type of metadata that can be stored compared to traditional relational databases.

In the sister eBank project, raw crystallographic data was annotated with metadata and "published" by being archived in the UK National Data Store as a "Crystallographic e-Print". Publications can then be linked back directly to the raw data for other researchers to access and analyze or verify. Another noteworthy feature of the project was that pervasive computing devices were used to capture laboratory conditions so that chemists could be notified in real time about the progress of their experiment using hand held PDAs.

The imminent data deluge: a key driver for e-Science
One of the key drivers underpinning the e-Science movement is the imminent availability of large amounts of data arising from the new generations of scientific experiments and surveys (Hey and Trefethen, 2003). New high-throughput experimental devices are now being deployed in many fields of science – from astronomy to biology – and this will lead to a veritable deluge of scientific data over the next five years or so. In order to exploit and explore the many Petabytes of scientific data that will arise from such next-generation scientific experiments, from supercomputer simulations, from sensor networks and from satellite surveys, scientists will need the assistance of specialized search engines and powerful data mining tools. To create such tools, the primary data will need to be annotated with relevant metadata giving such information as to the provenance, content and the conditions that produced the data. Over the course of the next few years, scientists will create vast distributed digital repositories of scientific data that will require management services similar to those of more conventional digital libraries as well as other data-specific services. As we have stressed, the ability to search, access, move, manipulate and mine such data will be a central requirement – or a competitive advantage – for this new generation of collaborative data-centric e-Science applications.

With this imminent deluge of scientific data, the issue of how scientists can manage these vast datasets becomes of paramount importance. Up to now, scientists have generally been able to manually manage the process of examining the experimental data to identify potentially interesting features and discover significant relationships between them. In the future, when we consider the massive amounts of data being created by simulations, experiments and sensors, it is clear that in many fields they will no longer have this luxury. The discovery process – from data to information to knowledge – needs to be automated as far as possible. At the lowest level, this requires automation of data management with the storage and organization of digital entities. At the next level, we require automatic annotation of scientific data with metadata describing both interesting features of the data and of the storage and organization of the resulting information. Finally, we will need new tools to enable scientists to progress beyond the generation of mere structured information towards the automated knowledge management of our scientific data.

The future of scholarly communication

The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities was drafted in 2003 "to promote the internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider" (Berlin Declaration, 2003). Signatories to the original declaration included research organizations such as the Fraunhofer and Max Planck Institutes in Germany, Centre National de la Recherche Scientifique (CNRS) and Institut National de Recherche en Informatique et en Automatique (INRIA) in France, Royal Netherlands Academy of Arts and Sciences (KNAW) and SURF in The Netherlands, JISC in the UK, CERN and Swiss Federal Institute of Technology (ETH) in Switzerland as well as many other international organizations and universities. The Berlin meeting followed in the footsteps of the Budapest Open Access Initiative in 2001[4]. It is important to recognize that the Berlin Declaration is not just concerned with textual material. The declaration defines open access contributions to include "original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material" (Berlin Declaration, 2003).

The research community is responding to the challenge of open access in a number of ways. Consider the three leading "prophets" of open access – Paul Ginsparg of arXiv[5], David Lipman of PubMed Central[6] and Stevan Harnad of EPrints[7].

The theoretical particle physics community had long had a tradition of circulating hard copy preprints of papers submitted to conventional journals ahead of review and publication. In the fast moving field of theoretical physics, the community is used to discussing the latest ideas at informal seminars and workshops and it makes no sense to attempt to delay discussion until after formal publication. With such a well-established preprint tradition, it was a natural but very significant step for Paul Ginsparg to establish an electronic archive at Los Alamos, where e-prints, electronic versions of preprints, could be displayed on a web site. From these small beginnings, Ginsparg has demonstrated a new way of scholarly communication outside the traditional scholarly publishing route of refereed journal articles. The arXiv has now moved to Cornell where it is owned and managed by the Cornell library and

this is now the standard first port of call for scientists in several subfields of physics, mathematics, computer science and quantitative biology[5]. It is interesting that the arXiv has no formal refereeing process to restrict publication on the site. Perhaps it is the very mathematical nature of the field that prevents the site from being overwhelmed by "noise" of low quality material. This mode of publication leads to many headaches for librarians of course. The published journal version of the original e-print may have revisions to the text and will certainly have a different layout and pagination. Proliferation of versions – e-prints, preprints, postprints and so on – as well as confusion about the precise date of "publication" are all now areas of concern to librarians. From a scientific point of view, these issues may seem trivial – since there is no doubt that claims for priority would be determined by the date of the e-print – but they are not at all trivial from the perspective of librarians and archivists[8].

The National Institutes of Health (NIH) in the US has a mandate to make publicly available a National Library of Medicine (NLM) of biomedical and healthcare resources. Their Entrez Life Sciences search engine gives access to both the PubMed service containing over 16 million citations from the MEDLINE database and life science journals for biomedical articles going back to the 1950s as well as a wide collection of biological databases. In February 2005, the NIH announced a new policy designed to accelerate the public's access to published articles resulting from NIH-funded research. The policy calls on scientists to release to the public manuscripts from research supported by NIH as soon as possible, and within 12 months of final publication. These peer-reviewed, NIH-funded research publications are now available in PubMed Central (PMC), a web-based archive managed by the National Center for Biotechnology Information (NCBI) for the NLM[6]. The online archive will increase the public's access to health-related publications at a time when demand for such information is on a steady rise. In their announcement, NIH Director Elias A. Zerhouni, M.D. said:

> With the rapid growth in the public's use of the internet, NIH must take a leadership role in making available to the public the research that we support. While this new policy is voluntary, we are strongly encouraging all NIH-supported researchers to release their published manuscripts as soon as possible for the benefit of the public. Scientists have a right to see the results of their work disseminated as quickly and broadly as possible, and NIH is committed to helping our scientists exercise this right. We urge publishers to work closely with authors in implementing this policy (*NIH News*, 2005).

The NIH policy for PubMed Central has several important goals, including:

- creating a stable archive of peer-reviewed research publications resulting from NIH-funded studies to ensure the permanent preservation of these vital research findings;
- securing a searchable compendium of these research publications that NIH and its awardees can use to manage more efficiently and to understand better their research portfolios, monitor scientific productivity, and, ultimately, help set research priorities; and
- making published results of NIH-funded research more readily accessible to the public, health care providers, educators, and scientists.

Beginning May 2, 2005, the policy requests that NIH-funded scientists submit an electronic version of the author's final manuscript, upon acceptance for publication, resulting from research supported in whole or in part by NIH. The author's final manuscript is defined as the final version accepted for journal publication, and includes all modifications from the publishing peer review process. The present policy gives authors the flexibility to designate a specific time frame for public release – ranging from immediate public access after final publication to a 12 month delay – when they submit their manuscripts to NIH. Authors are strongly encouraged to exercise their right to specify that their articles will be publicly available through PMC as soon as possible. With the addition of PubMed Central, Entrez searches can now be directed to free full text versions of the research article.

Jim Gray and Jean Paoli from Microsoft have worked with David Lipman and the NCBI team to develop a "portable" version of PubMed Central which is now being deployed in other countries around the world. The NLM's archiving template for XML documents – the Document Type Definition or DTD – is now becoming the international standard for such archives. The Wellcome Trust in the UK, in partnership with the JISC and the NLM are working together on a project to digitize the complete backfiles of a number of important and historically significant medical journals[9]. The digitized content will be made freely available on the internet – via PMC – and augment the content already available there. The Wellcome Library exists as a resource to provide access to the documentary record of medicine. This project is one way of translating that vision into the digital age.

The two repositories described above are examples of subject specific repositories. By contrast, Stevan Harnad advocates author "self-archiving" in departmental or institutional repositories (Harnad and Hey, 1995). Open access archives or repositories are digital collections of research articles that have been placed there by their authors. In the case of journal articles this may be done either before (preprints) or after publication (postprints). These repositories expose the metadata of each article (the title, authors, and other bibliographic details) in a format compliant with the open archives initiative protocol for metadata harvesting (OAI-PMH)[10]. As a result, OAI-compliant search engines can harvest the metadata from each repository into large databases of worldwide research, which researchers can then use to locate articles of interest. Open access repositories can be centralized and subject-based, such as arXiv and PubMed Central, or they may be distributed and multidisciplinary, located in universities or other research-based institutions. A list of open access archives is maintained at the Registry of Open Access Repositories (ROAR)[11] and OpenDOAR sites[12].

From these examples, one sees that the nature of scholarly publishing is changing. Not only is publication on the web, in one form or other, enabling access to a much wider range of research literature but also we are seeing the emergence of data archives as a complementary form of scholarly communication. In some fields, such as biology, databases are already one of the primary mechanisms of scholarly publishing. In the area of environmental science, the Natural Environment Research Council, UK (NERC) DataGrid project is building a grid that makes data discovery, delivery and use much easier than it is now, facilitating better use of the existing investment in the curation and maintenance of quality data archives[13]. This DataGrid project also intends to make the connection between data held in managed archives and data held by

individual research groups seamless in such a way that the same tools can be used to compare and manipulate data from both sources. When fully functional, it will deliver scientists the completely new ability of being able to compare and contrast data from an extensive range of US and European datasets from within one specific context. What is the role of the library community in this new world of scientific information management? One relevant example is that of the SPIRES Digital Library[14]. The SPIRES-HEP database has been run by the Stanford Linear Accelerator Center (SLAC) since the late 1960s as a database of particle physics literature. As mentioned in the introduction, this became the first web-site in North America in 1991 and now attracts around 50,000 searches per day from particle physicists. In contrast to just accessing research literature via arXiv, SPIRES offers access to the whole of the HEP literature with arXiv as only one of its key resources. The database is managed and maintained by the SLAC Library, in cooperation with librarians and particle physicists from DESY, FNAL, Kyoto, Durham, IHEP and KEK.

The above examples concern national and international data archives. However there is also likely to be a role for libraries at the institutional level in curating and preserving e-Science data in addition to their more traditional role in organizing and curating digital research output in the form of journal papers, reports and theses. Consider the following quotation from the rationale for MIT's DSpace repository:

> Much of the material produced by faculty, such as datasets, experimental results and rich media data as well as more conventional document-based material (e.g. articles and reports) is housed on an individual's hard drive or department Web server. Such material is often lost forever as faculty and departments change over time (Tansley *et al.*, 2003, p. 87).

Since some of this data may be relevant for the protection of the university's intellectual property, it is obvious that universities and libraries need to be clear about their roles in the curation and preservation of such data. In the next section, we describe a case study from the University of Southampton in the UK where there is experimentation with some of the different roles for the library in supporting research at the university.

An institutional repository at Southampton: The TARDis experience

Traditionally, academic libraries have played a major role in undergraduate teaching and information retrieval skills. The large increases in student numbers and pace of change in the web based environment have caused librarians to constantly reexamine their skills and tasks for their support of undergraduate education. By contrast, research support has usually been concentrated on providing access to research resources external to the university and assisting the researcher in accessing original sources, e.g. by access to electronic versions of journals or to hard copy versions of papers and books via inter library loans. However, the growing emphasis on capturing research outputs at an earlier stage in the scholarly communication cycle now provides a significant opportunity for librarians to rethink their role as information managers, strategists and custodians of the research environment.

The JISC funded Focus on Access to Institutional Resources (FAIR) program in the UK gave an opportunity to the Targeting Academic Research for Deposit and Disclosure (TARDis) project at the University of Southampton to investigate the practical implications of creating an institutional repository for research and to explore

the relationship between an institutional repository and an open access archive (Simpson and Hey, 2005). The project undertook a survey of the attitudes of the university's researchers – from senior management to individual academics – and their key feedback was to underline the importance of integrating a repository both into the university's current research management needs and also integrating the deposit process into the researcher's work practice. In the case of Southampton, a crucial functionality of the institutional repository was the ability to record publications for use by the university, by the department, by the research group and by individuals at an early stage in the scholarly research cycle – rather than at some more remote time such as that corresponding to formal publication which can be long after the initial production of a research output. Information capture could therefore take place either at the working paper stage or at the more final published paper and book chapter stage. Such a repository must empower academics to manage their own information management demands more efficiently and enable them to immediately add the full text or have the option to add this later if this is a more practical solution. A good summary of the goals of an Institutional Repository has been given by Lynch (2003).

The TARDis project was also able to feed information management requirements to the developers of the EPrints software at Southampton. In particular, they were able to influence the provision of fields and citation styles necessary to allow flexible reuse of the metadata. For example, in an institutional context, setting up a separate database solely for papers available with full text would require a huge duplication of effort if implemented on a university scale. The TARDis model therefore simply requires that searches of the whole database should reflect all types of research output and that searches for "full text only" items can be obtained from the "Open Access Archive" – the subset of research outputs for which the full text is stored on the same server. In the future, with changing attitudes to open access globally and with researchers becoming more familiar with saving and depositing their full text, the TARDis route map shows that the Institutional Repository comes closer to the vision of the Open Access archive (Hey *et al.*, 2005). However, with the increase in content in repositories such as arXiv and PubMed Central and other subject or conference based archives it seems increasingly likely that the research repository, as it grows in size and complexity, will be a pragmatic mix of full text, where the process of deposition is either straightforward or where there is a need to ensure there is a local copy, and of links to trusted repositories where this is more practical. By this time, it will be more meaningful for researchers to search the whole institutional repository rather than just the subsection of the archive that stores the full text locally. These are the kinds of information management decisions that librarians will have to make in the future just as they have traditionally weighed up whether to buy or acquire items on interlibrary loan. This mixed economy for the institutional repository has analogies with both the SPIRES database and its relationship with arXiv, other full text repositories and journal sites and with the Entrez Search Engine, PubMed and the free full text PubMed Central repository.

The university publications database must portray the full picture of all research outputs: this is key to the goal of representing all disciplines fairly – not just those that follow the traditional peer reviewed scientific journal model. Listening to feedback from all parts of the academic community at Southampton has therefore resulted in a

more complex project than had originally been envisioned by the library – namely, that of creating a full publications database rather than just a digital repository. However, this enlargement of scope has enabled the project to move from being just a pilot project to one that is now seen as an integral part of the university's research management infrastructure – and one that is able to respond to demands for open access on a more gradual but more sustainable timescale.

Where a national publications recording system is already in place such as in The Netherlands or in Australia, other information management decisions may need to be made. For example, there may be a need for practical steps to be taken by librarians to simplify recording in both a publications database and a full text database in an efficient manner (Woodland and Ng, 2006). In either case it is important for the original author to be part of the process to ensure the full text is deposited where possible and for the author to be aware of the potential for easy reuse of the metadata in CVs, project reports and proposals and for numerous other publication management demands.

The TARDis project was focused on research output but it is possible to envision a more ambitious role for an institutional repository as that of embracing the entire intellectual output of an institution. In working towards such a goal, there is much that the library can learn from the infrastructure required for a research library – both in recording research outputs and in the management of both publications and data. For example, the National Oceanography Centre at Southampton (NOC,S) is one of the world's leading centres for research and education in marine and earth sciences, for the development of marine technology and for the provision of large scale infrastructure and support for the marine research community. The National Oceanographic Library at NOC,S has long had a traditional role in recording research publications but also played a major role in the TARDis project and in the development of the University of Southampton Research Repository. It is now investigating the role of the library in the management and preservation of local data sources. Through the JISC funded Citation, Location, and Deposition in Discipline and Institutional Repositories (CLADDIER) project, the National Oceanographic Library is exploring the linking of its publications in the Institutional Repository with environmental data holdings[15]. The result will be a step on the road to a situation where active environmental scientists will to be able to move seamlessly from information discovery (location), through acquisition to deposition of new material, with all the digital objects correctly identified and cited. Experience at Southampton shows that a partnership between librarians and researchers is likely to give the best results – an experienced information manager/librarian is helpful in creating good citations for data entities (now given unique digital object identifiers – DOIs) in the repository. Another example of the need for links between the Southampton Institutional Repository and a data archive is that of the eCrystals Crystal Structure Report Archive[16]. Southampton is the home of the National Archive for Crystal Structures generated both by the Southampton Chemical Crystallography Group and by the Engineering and Physical Sciences Research Council (EPSRC) UK National Crystallography Service that is located on the Southampton campus. This raises questions as to which organization owns the long-term responsibility for a national service that is created from a project by academics at the institution. In one model, this can be seen as just another strand of the institutional repository in capturing all intellectual assets – publications, data, learning objects and outputs such as proceedings and papers from workshops. The

lessons learned from these examples will be valuable in establishing clear relationships and responsibilities between discipline based repositories and institutional repositories.

There are many other issues – such as those of provenance and preservation. In many research fields there are national repositories responsible for the curation and preservation of their scientific data. University libraries, on the other hand, may need to take responsibility for assisting with the curation and preservation of smaller scale data sets arising from the research of research groups or individual academics. The increasing importance of digital Institutional Repositories is giving an impetus to examine the associated preservation issues (Hitchcock *et al.*, 2005). Repository administrators will need to be supported in these underlying issues so that they can concentrate on their key goals of recording and providing access to scholarly output.

Conclusions

The advent of e-Science heralds a new and exciting world for the library world to be involved in. In both Europe and the USA there are now moves to develop a powerful infrastructure to support collaborative, multidisciplinary science. Such infrastructure is termed "e-Infrastructure" in Europe and Cyberinfrastructure in the US. One component of this infrastructure will be "grid" middleware that enables researchers to easily set up their own secure "virtual organizations" linking research sites with whom they wish to share a variety of resources with controlled authenticated access. A second ingredient of this research infrastructure is of course the underlying research network that constitutes the academic research Internet. The last key ingredient of the research infrastructure is access to research results – both publications and data. Thus the e-Science revolution will put libraries and repositories center stage in the development of the next generation research infrastructure.

Notes

1. www.CombeChem.org
2. www.ukoln.ac.uk/projects/ebank-uk
3. www.SmartTea.org
4. www.soros.org/openaccess/
5. http://arxiv.org/
6. www.pubmedcentral.nih.gov
7. www.eprints.org/
8. The VERSIONS project: http://library-2.lse.ac.uk/versions/
9. The Wellcome Library: http://library.wellcome.ac.uk/node280.html
10. www.openarchives.org/OAI/openarchivesprotocol.html
11. http://archives.eprints.org/
12. www.opendoar.org/
13. http://ndg.nerc.ac.uk
14. www.slac.stanford.edu/spires/
15. http://claddier.badc.ac.uk
16. http://ecrystals.chem.soton.ac.uk/

References

Atkins, D., Droegemeier, K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., Messina, P., Ostriker, J.P. and Wright, M.H. (2003), *Revolutionizing Science and Engineering through Cyberinfrastructure*, report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, available at: www.nsf.gov/cise/sci/reports/atkins.pdf (accessed April 4, 2006).

Berlin Declaration (2003), "Berlin declaration on open access to knowledge in the sciences and humanities", available at: www.zim.mpg.de/openaccess-berlin/berlindeclaration.html (accessed April 4, 2006).

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), "The semantic web", *Scientific American*, Vol. 284 No. 5, pp. 34-43.

Frey, J.G., De Roure, D. and Carr, L.A. (2002), "Publication at source: scientific communication from a publication web to a data grid", paper presented at the Euroweb 2002 Conference: The Web and the Grid: From e-Science to e-Business, Oxford, December 17-18, available at: http://ewic.bcs.org/conferences/2002/euroweb/session3/paper3.pdf (accessed April 4, 2006).

Frey, J.G., Bradley, M., Essex, J.W., Hursthouse, M.B., Lewis, S.M., Luck, M.M., Moreau, L., De Roure, D., Surridge, M. and Welsh, A. (2003), "Combinatorial chemistry and the grid", in Berman, F., Fox, G. and Hey, T. (Eds), *Grid Computing: Making the Global Infrastructure a Reality*, Wiley, Chichester, pp. 945-62.

Friedman, T. (2005), *The World is Flat: A Brief History of the Twenty-First Century*, Farrar, Strauss and Giroux, New York, NY.

Harnad, S. and Hey, J. (1995), "Esoteric knowledge: the scholar and scholarly publishing on the net", in Dempsey, L., Law, D. and Mowat, I. (Eds), *Networking and the Future of Libraries 2: Managing the Intellectual Record*, Library Association Publishing, London, pp. 110-6.

Hey, J., Simpson, P. and Carr, L. (2005), "The TARDis route map to open access: developing an institutional repository model", in Dobreva, M. and Engelen, J. (Eds.), *ELPUB2005 From Author to Reader: Challenges for the Digital Content Chain: Proceedings of the 9th ICCC International Conference on Electronic Publishing, Katholieke Universiteit Leuven, Leuven-Heverlee, Belgium, 8-10 June 2005*, Peeters Publishing, Leuven, pp. 179-82, available at: http://eprints.soton.ac.uk/16262/ (accessed April 4, 2006).

Hey, T. and Trefethen, A. (2003), "The data deluge", in Berman, F., Fox, G. and Hey, T. (Eds), *Grid Computing: Making the Global Infrastructure a Reality*, Wiley, Chichester, pp. 809-24.

Hitchcock, S., Brody, T., Hey, J. and Carr, L. (2005), "Preservation for institutional repositories: practical and invisible", in *Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data (PV2005)*, The Royal Society, Edinburgh, 21-23 November 2005, available at: http://eprints.soton.ac.uk/18774/ (accessed April 4, 2006).

Hughes, G., Mills, H., De Roure, D., Frey, J.G., Moreau, L., Schraefel, M.C., Smith, G. and Zaluska, E. (2004), "The semantic smart laboratory: a system for supporting the chemical e-Scientist", *Organic and Biomolecular Chemistry*, Vol. 2 No. 22, pp. 3284-93.

Lynch, C. (2003), "Institutional repositories: essential infrastructure for scholarship in the digital age", *ARL BiMonthly Report*, No. 226, pp. 1-7, available at: www.arl.org/newsltr/226/ir.html (accessed April 4, 2006).

NIH News (2005), "NIH calls on scientists to speed public release of research publications", February 3, available at: www.nih.gov/news/pr/feb2005/od-03.htm (accessed April 4, 2006).

Schraefel, M.C., Hughes, G., Mills, H., Smith, G. and Frey, J.G. (2004a), "Making tea: iterative design through analogy", in *Proceedings of the Conference on Designing Interactive Systems*, available at: http://eprints.soton.ac.uk/15879/01/schraefel-MakingTeaDIS04.pdf (accessed April 4, 2006).

Schraefel, M.C., Hughes, G., Mills, H., Smith, G., Payne, T. and Frey, J. (2004b), "Breaking the book: translating the chemistry lab book to a pervasive computing environment", in *Proceedings of the Conference on Human Factors (CHI)*, available at: http://eprints. soton.ac.uk/15878/01/schraefel_smartteaChi04.pdf (accessed April 4, 2006).

Segaller, S. (1998), *Nerds: A Brief History of the Internet*, TV Books, New York, NY.

Simpson, P. and Hey, J. (2005), "Institutional e-Print repositories for research visibility", in Drake, M. (Ed.), *Encyclopedia of Library and Information Science*, 2nd ed., Dekker, New York, NY, available at: http://eprints.soton.ac.uk/9057/ (accessed April 4, 2006).

Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G. and Smith, M. (2003), "The DSpace Institutional Digital Repository System: current functionality", in *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'03)*, pp. 87-97.

Taylor, J.M. (2001), Talk given at UK e-Science Town Meeting, July.

Woodland, J. and Ng, J. (2006), "Too many systems, too little time: integrating an eprint repository into a university publications system", in *VALA 2006 13th Biennial Conference and Exhibition, Crown Towers Melbourne, Australia, 8-10 February 2006*, Victorian Association for Library Automation, Melbourne, available at: http://espace.lis. curtin.edu.au/archive/00000618/ (accessed April 4, 2006).

Corresponding author

Tony Hey can be contacted at: tonyhey@microsoft.com

Exploring the academic invisible web

Dirk Lewandowski

*Department of Information Science, Heinrich-Heine-University,
Düsseldorf, Germany, and*

Philipp Mayr

*Institute of Library and Information Science, Humboldt-University,
Berlin, Germany*

Abstract

Purpose – The purpose of this article is to provide a critical review of Bergman's study on the deep web. In addition, this study brings a new concept into the discussion, the academic invisible web (AIW). The paper defines the academic invisible web as consisting of all databases and collections relevant to academia but not searchable by the general-purpose internet search engines. Indexing this part of the invisible web is central to scientific search engines. This paper provides an overview of approaches followed thus far.

Design/methodology/approach – Provides a discussion of measures and calculations, estimation based on informetric laws. Also gives a literature review on approaches for uncovering information from the invisible web.

Findings – Bergman's size estimate of the invisible web is highly questionable. This paper demonstrates some major errors in the conceptual design of the Bergman paper. A new (raw) size estimate is given.

Research limitations/implications – The precision of this estimate is limited due to a small sample size and lack of reliable data.

Practical implications – This study can show that no single library alone will be able to index the academic invisible web. The study suggests a collaboration to accomplish this task.

Originality/value – Provides library managers and those interested in developing academic search engines with data on the size and attributes of the academic invisible web.

Keywords Search engines, Worldwide Web, Indexing, Digital libraries

Paper type Research paper

Introduction

Recent years demonstrate an unbroken trend towards end-user searching. Users expect search services to be complete, integrated and up-to-date. Educated users naturally want to retrieve the most comprehensive and largest index. But size is not the only issue. Even in the academic sector, where advanced search tools and dozens of relevant reference and full text databases are to be found, users to a large degree consult general-purpose internet search engines to retrieve academic documents. Information professionals, who are used to tackling multiple data sources and varied, combined search environments, are forced to use oversimplified, general search engines.

The rise of web search engines has brought with it some shifts in user behavior. Web search engines suggest that all information available can be searched within just one system. The search process itself is easy and highly self-explanatory. Within the

last few years, professional information vendors (and libraries) have found that search engine technology can easily fit their needs for making academic content available for end-user searching. Keeping in mind that search engine technology is also widely used in a business context, it can be said that this technology is the new key concept in searching (see Lewandowski, 2006).

The reasons for this shift in information behavior are relatively clear. More and more scholarly content is provided exclusively on the web. The open access movement is only one current example for this paradigm change: from the traditional print publishing system to the electronic publishing paradigm. The consequence is a situation that Krause calls the poly-central information provision (Krause, 2003). A growing decentralization in the field of new information providers and changed user expectations and habits have led to a gap in the providing of information. General search engines take advantage of this gap. Google Scholar and Scirus show this very clearly: they do index parts of the invisible web, but unfortunately with results of questionable quality (see below). A recent review of existing technologies to index the invisible web can be found in Ru and Horowitz (2005). They identified the main problems and strategies in indexing the invisible web. According to Ru and Horowitz "indexing the web site interface" or "examining a portion of the contents" of an invisible web site are the two typical approaches.

The pivotal point in the dilemma is the invisible web (for a detailed discussion see Lewandowski, 2005b). Library collections and databases with millions of documents remain invisible to the eyes of users of general internet search engines. Furthermore, ongoing digitization projects are contributing to the continuous growth of the invisible web. Extant technical standards like Z39.50 or Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) are often not fully utilized, and consequently, valuable openly accessible collections, especially from libraries, remain invisible. It could be asked whether general-purpose search engines should pay more attention to the invisible web, but, as has been demonstrated in recent years, these seem to be lacking in terms of completeness and information quality (see Mayr and Walter, 2005; Brophy and Bawden, 2005). So other institutions with experience in information organization should attend to this task.

The structure of this article is as follows: first, we discuss the competing definitions of the invisible web and give a definition for the academic invisible web. Then, we retrace Bergman's study on the size of the invisible web, in which we find some serious errors. We suggest new approaches to determine a better size estimate. In the next part of the article, we discuss the approaches used so far to uncover information from the invisible web. In the discussion section, we offer implications as to how libraries should deal with the issue of the academic invisible web and give a roadmap for further research on the topic.

Defining the (academic) invisible web

In short, the invisible web is the part of the web that search engines do not add to their indices. There are several reasons for this, mainly limited storage space and the inability to index certain kinds of content. We discuss two definitions of the invisible web, where we do not distinguish between the invisible web and the deep web. Both terms are widely used for the same concept and using one or the other is just a matter of preference. We use the established term invisible web. Sherman and Price give the following definition for the invisible web:

> Text pages, files, or other often high-quality authoritative information available via the worldwide web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of web pages (Sherman and Price, 2001, p. 57).

This is a relatively wide definition as it takes into account all file types and includes the inability of search engines to index certain content as well as their choice not to index certain types of content. In this definition, for example, spam pages are part of the invisible web because search engines choose not to add them to their indices.

Bergman defines this much more narrowly. Focusing on databases available via the web, he writes:

> Traditional search engines cannot "see" or retrieve content in the deep web – those pages do not exist until they are created dynamically as the result of a specific search (Bergman, 2001).

Table I shows the different types of invisible web content according to Sherman and Price. It is easy to see that their view of the invisible web includes Bergman's view in the rows "content of relational databases" and "dynamically generated content."

Disconnected pages are a real problem of the invisible web, but to a lesser extent than with the surface web. If search engines could find these pages, there would be no problem indexing them. There is the technical problem of a lack of information about the existence of these pages.

Some other, more technical problems, such as dynamically generated pages and file types, have nearly been solved by now. It remains true that programs and compressed files are not readable for search engines, but this begs the question of what is the use of search engines being able to index these. Other file types mentioned by Sherman and Price, such as PDF, are read by all major search engines nowadays. But Flash and Shockwave content still remain a problem, due to the lack of sufficient text for the search engines to index. The main problem here lies in the inability of most search engines to follow links within flash sites.

Real-time content remains a problem because search engines cannot keep up with the rapid update rates of some sites. But in the current context of indexing the

Type of invisible web content	Why it's invisible
Disconnected page	No links for crawlers to find the page
Page consisting primarily of images, audio, or video	Insufficient text for the search engine to "understand" what the page is about
Pages consisting primarily of PDF or Postscript, Flash, Shockwave, executables (programs) or compressed files (.zip,.tar, etc.)	Technically indexable, but usually ignored, primarily for business or policy reasons
Content in relational databases	Crawlers cannot fill out required fields in interactive forms
Real-time content	Ephemeral data; huge quantities; rapidly changing information
Dynamically generated content	Customized content is irrelevant for most searchers; fear of "spider traps"

Source: Sherman and Price (2001, p. 61)

Table I.
Types of invisible web content

academic invisible web, this content type can be left out. This also holds true for the other technical limitations described by Sherman and Price. Therefore, we think that efforts in indexing the invisible web in general, and the academic part of it in particular, should primarily focus on databases not visible to general search engines. Therefore, we stick to Bergman's definition of the invisible web. Particularly in the academic context, the content of databases is central. Technical limitations do not need to be taken into consideration for academic content, because it is mainly in formats such as PDF, which are technically readable by general-purpose search engines.

But not all limitations in indexing the invisible web are purely technical. Sherman and Price define four types of invisibility, where, for our purposes, the distinction between proprietary and free content is important. A large part of the invisible web relevant to academia is part of the proprietary web, mainly the content from publishers' databases.

From a library perspective, the academic invisible web consists mainly of text documents (in different formats such as PDF, PPT, DOC). This is the content that libraries (or academic search engines) should add to their searchable databases to give the user a central access point to all relevant content.

Therefore, we define the academic invisible web (AIW) as consisting of all databases and collections relevant to academia but not searchable by the general internet search engines.

In accordance with Lossau's claim that libraries need to discover the academic internet (Lossau, 2004), one could narrow the above definition to the content of the databases that should be indexed by libraries (using search engine technology). We do not intend to say that one library alone should make all content from the AIW visible in a search engine, but that libraries should follow a cooperative approach in making this content visible.

It should be kept in mind that the AIW is only one part of the web relevant to libraries. The academic surface web (ASW) contains a multitude of relevant documents as well, e.g. most open access repositories are part of the surface web and can be crawled by general-purpose search engines without any problem. The study by Lawrence and Giles (1999) returned results showing that only about 6 percent of the indexable web are academic content.

The AIW is valuable for scholars, librarians, information professionals and all other academic searchers and can provide everything relevant to the scientific process. This includes:

- literature (e.g. articles, dissertations, reports, books);
- data (e.g. survey data); and
- pure online content (e.g. open access documents).

The main institutional providers of AIW content are:

- database vendors, producing bibliographic metadata records enriched by human subject indexing (thesauri, classifications and other knowledge organization systems) and additional services like document delivery;
- libraries, also producing bibliographic records in openly accessible systems like online public access catalogues (OPACs), offering their collections enriched by human subject indexing and additional services;
- commercial publishers, providing mainly full text content;

- other repositories of societies and corporations (e.g. the Association for Computing Machinery); and
- open access repositories (e.g. Citebase, OpenROAR).

A lot of these materials are not necessarily part of the AIW, but are in fact uncovered by the main search engines and tools. For users of these heterogeneous collections, this means becoming accustomed to the respective systems and information structures. For example, most providers of scholarly information maintain their own subject access and information organization models, due to various traditions and indexed content types. Libraries index mainly books and compilations with their standardized universal authority files; database producers use proprietary domain-specific thesauri and classifications for indexing journal articles, while publishers use a mixture of manual and automatic indexing for their full texts. This results in a heterogeneity (Krause, 2003) between the collections and a complex situation for users in need of cross-database searching.

Measuring the size of the (academic) invisible web

To our knowledge, the only attempt to measure the size of the invisible web was Bergman's (2001) study. The main findings were that the invisible web is about 550 times larger than the surface web and consists of approximately 550 billion documents. Bergman's paper is widely cited and therefore we will discuss it in detail. Most other studies use Bergman's size estimates or estimate the size of the invisible web based on the ratio between surface and invisible web of 1:550 given by Bergman (e.g. Lyman *et al.*, 2003).

The basis for Bergman's size estimates is a "top 60" list containing the largest deep web sites. These are put together manually from directories of such sites, while duplicates are removed. Bergman's top 60 contains 85 billion documents with a total size of 748,504 GB. The top two alone contain 585,400 GB, which is more than 75 percent of the top 60 (file size measure).

A further assumption is that there are around 100,000 deep web databases. This number comes from an overlap analysis between the largest directories of invisible web sites. Bergman's further calculations use the mean size of 5.43 million documents per invisible web database. Therefore, he states that the total size of the invisible web (mean multiplied by the number of databases) is 543 billion documents. Bearing in mind that the size of the surface web at the time of the investigation (2001), was approximately 1 billion documents (based on data from Lawrence and Giles, 1999), Bergman finds that the invisible web is 550 times larger than the surface web.

These numbers were soon challenged (Sherman, 2001; Stock, 2003), but these authors just made new guesses and did not deliver a new calculation or even an explanation as to why Bergman's figures had to be mistaken. Our investigation found that the error lies in the use of the mean for the calculation of the total size estimate. While the mean is very high, the median of all databases is relatively low with just 4,950 documents. Looking at Bergman's top 60 list, we see that the distribution of database sizes is highly skewed (Figure 1), so the mean cannot be used to calculate the total size.

The skewed distribution of database sizes is typical and can also be seen in other database portfolios such as the DIALOG databases accessible via the web. Again we see a highly skewed distribution (long tail). The sizes of the 347 files in DIALOG are
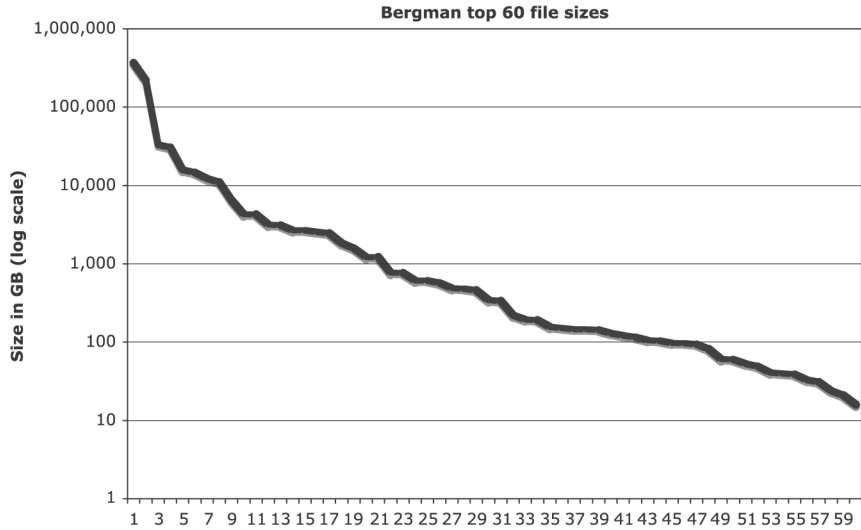
Figure 1.
Distribution of file sizes in
Bergman's top 60

plotted along a logarithmic scale (see Figure 2), demonstrating that there are few databases with more than 100,000,000 records (compare to Williams, 2005), and the majority with less than 1,000,000 records. The distribution is described by an exponential function with a high Pearson correlation (Pearson is 0.96, see Figure 2). The median of all 347 database sizes is circa 380,000 records. We hypothesize that the AIW will also follow such an exponential distribution.
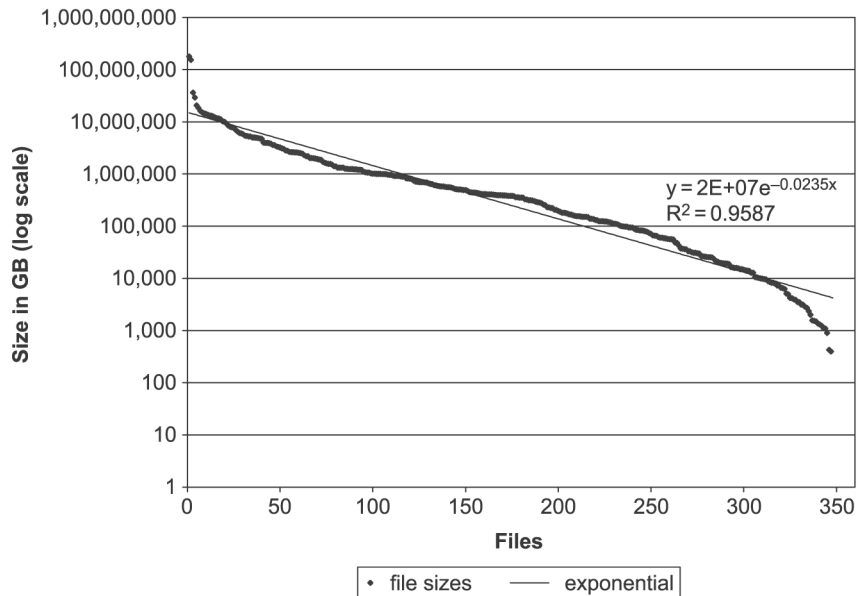

Figure 2.
Distribution of database
sizes from the DIALOG
host ($n = 347$)

For all further calculations in Bergman's study, the size in GB is used instead of the number of records per database. This is very problematic, as it is impossible to derive the record counts from the file size data due to the greatly varying size of database records (pictures, bibliographic records, full text records). Therefore, we are not able to make a more accurate calculation from Bergman's data. We can say that his size estimates are far too high, because of two fundamental errors. Firstly, the statistical error of using the mean instead of the median calculation, and secondly his misleading projection from the database size in GB. When using the 85 billion documents from his top 60, we can assume that the total number of documents will not exceed 100 billion because of the highly skewed distribution. Even though this estimate is based on data from 2001, we think that the typical growth rate of database sizes (cf. Williams, 2005) will not affect the total size to a large extent.

But how much of the invisible web is academic content? Looking at Bergman's top 60, we find that indeed 90 percent can be regarded as academic content, but if one chooses to omit all databases containing mere raw data, the portion of academic content shrinks to approximately four percent (Figure 3), which corresponds to the amount of academic content found on the surface web in the 1999 study from Lawrence and Giles. The main part of Bergman's invisible web consists of raw data, mainly pictures such as satellite images of the Earth. The records of these databases are far bigger than those of textual databases. Because Bergman only uses GB sizes, one cannot calculate new size estimates based on record numbers from the given data. For this task, one needs to build a new collection of the biggest invisible web databases.

In summary, Bergman's study exhibits shortcomings in the mixture of database types and database content, as well as the calculation method used. It goes beyond the

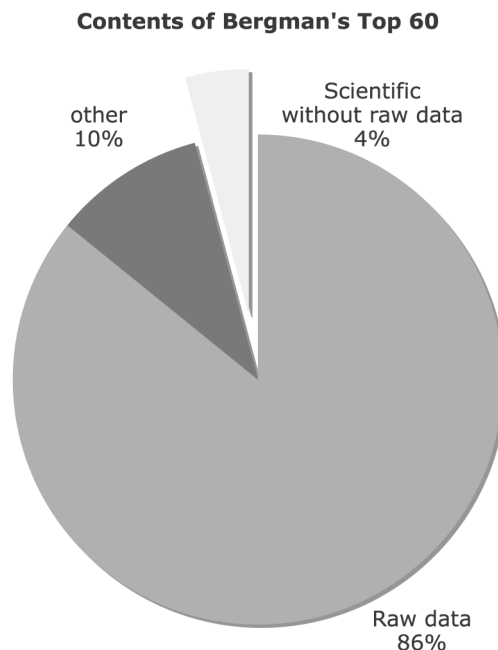**Contents of Bergman's Top 60**

Figure 3.
Contents of Bergman's
top 60

scope of this paper to present an exact size estimate for the invisible web. Further research is needed. In particular, a reliable collection of the largest invisible web databases should be built.

As we are not satisfied with Bergman's size estimates or our own raw estimate from Bergman's data, we have used additional data from the Gale Directory of Databases (Williams, 2005) for comparison. The directory contains approximately 13,000 databases and covers all major academic databases, as well as a number of databases solely of commercial interest. The total size estimate for all databases is 18.92 billion documents. The average size per database is 1.15 million records, with a highly skewed distribution. A total of 5 percent of the databases contain more than one million records, some more than 100 million. Omitting these very large databases, the mean database size is about 150,000 records. The total size estimate is calculated by adding the known database sizes and assuming the mean of 150,000 records for every other database. This method only works when all the very large database sizes are known. We cannot verify if all these are considered, but we found that some of the databases included in Bergman's top 60 are missing from the Gale Directory. Therefore, it is difficult to compare the numbers directly. Because of the missing databases, the numbers from Gale are probably too low. In conclusion, we can only make an educated guess as to the actual size of the AIW: in our opinion, its size lies between 20 and 100 billion documents, viewing the raw data as part of the AIW. If these data were to be omitted, the AIW would be far smaller. What we can definitely say is that the size of the AIW lies within the range of the index sizes of the biggest surface web search engines (Lewandowski, 2005c). Therefore, the challenge in indexing the whole AIW can only be met through a cooperative effort and not by a single institution acting alone.

Williams (2005) divides the databases in the Gale directory into six classes: word-oriented, number-oriented, image/video, audio, electronic services and software. For libraries and academic search engines, it is mainly word-oriented databases, comprising about 69 percent of all databases, which are of interest. Of these 8,994 word-oriented databases, some 80 percent are full-text or bibliographic information. We feel that these numbers represent a good starting point when attempting to index the whole academic invisible web.

Approaches to indexing and opening the academic invisible web

There are different models for enhancing access to the AIW, of which we can mention only a few. The four systems to be described shortly have a common focus on scholarly information, but the approaches and the content they provide are largely different. Google Scholar and Scirus are projects started by commercial companies. The core of their content is based on publishers' repositories plus openly accessible materials. On the other hand, Bielefeld Academic Search Engine (BASE) and Vascoda are academic projects where libraries and information providers open their collections, mainly academic reference databases, library catalogues plus free extra documents (e.g. surface web content). All systems use or will use search engine technology enhanced with their own implementations (e.g. citation indexing, specific filtering or semantic heterogeneity treatment).

Google Scholar[1] is currently the most discussed approach (Notess, 2005). The beta version, online since November 2004, covers some million documents. Google Scholar indexes a substantial part of international science – technology – medicine (STM)

publishers and other publishers who joined from the Crossref initiative. Google set up a prototype with great potential, but which also exhibits some unwelcome characteristics (Lewandowski, 2005a; Mayr and Walter, 2005). To its credit, Google Scholar tries to adopt the influential citation measure introduced by the Institute of Scientific Information (ISI) and implemented in the former Science Citation Index, now Web of Science. Unfortunately, Google Scholar provides no documentation that would make the service more transparent (Jacsó, 2005). It is impossible to say anything about the exact coverage, or how up-to-date the current service is, as a recent empirical study shows (Mayr and Walter, 2005).

Scirus[2] (see Scirus, 2004) is a scientific search engine that indexes the academic surface web and also several other collections such as Elsevier's Science Direct and open access sources. This approach comes close to the desired combination of surface web content and AIW content, but is far from being complete, at least in the AIW part. With approximately 250 million documents from the surface web, Scirus is by far the largest search engine of its kind built with FAST technology (McKiernan, 2005).

BASE[3] (see Lossau, 2004) is an integrated search engine combining data from the library catalogue of Bielefeld University Library and data from approximately 160 open access sources (more than 2 million documents). It uses the FAST search engine.

Vascoda[4] is the prototype of an interdisciplinary science portal integrating library collections, literature databases and additional scholarly content. Vascoda acts as a meta portal delegating requests to lower, domain-specific layers or clusters. Each domain is responsible for its own subject portal that can be built using various technologies. Vascoda is an alternative model for a system bridging the gap of the AIW, designed by German libraries and documentation centers. Vascoda will soon launch its latest version enhanced by FAST search engine technology.

The roundup of these prototypical academic search systems shows clearly that serious efforts to index the AIW will need a collaborative approach. Every single approach has its own specific strengths and weaknesses. On the one hand we see broad cover-age with a bias towards commercial hits and the inability to exclude non-academic records from the results. On the other hand, we have limited scope and a lack of full text information.

Discussion and conclusion

Search engines are increasingly acquiring a gatekeeper function and are widely seen as offering general access to information due to their simplicity, search velocity and broad coverage. But this is true only for a part of the web.

As called for by Lossau (2004), libraries should discover the academic web. Although we focused on the academic invisible web, there are also parts of the visible web relevant to libraries. The key in achieving the best experience for the library user lies in a combined approach for both types of content. We were able to show that the AIW is very large and that its size is comparable to the indices of the largest general-purpose web search engines. Therefore, only a co-operative approach is possible.

We conclude that existing search tools and approaches show potential to make the AIW visible. What we do not see is a real will for lasting collaboration among the players mentioned. Commercial search engine providers with their technological and financial superiority should work together with libraries, which have long experience in collection building and subject access models. They developed complex instruments

for information organization (e.g. thesauri, classification, taxonomies) that could be highly valuable for end-user searching, automatic indexing, ontology building and classification of academic content. Publishers and database vendors should join by opening their collections (see Google Scholar example).

Unfortunately, we were not able to give a more precise size estimate for the academic invisible web. Further research should focus on this task. We need to build a collection of the largest AIW databases and use the informetric distribution that we assume to be also given for the AIW. A good size estimate could be given based on such a sample.

Another task is to classify the AIW content to get a picture of the extent to which the different disciplines contribute to its size. Recommendations as to how to build specialized search engines for the various disciplines could be given based on such a classification.

A final research task is the distinction between the visible and the invisible web. In the past years, we saw the conversion of large databases into HTML pages for the purpose of becoming indexed by the main web search engines. Although this is mainly done in the commercial context, some libraries followed this approach with varying degrees of success (see also Lewandowski, 2006). If database vendors make their databases available on the visible web, libraries could follow the approach of Google or other search engines in indexing this content. Further research on this topic is needed, because at the current time nobody knows to what extent database content is already available on the surface web.

We can further conclude that Bergman did a good job in bringing the topic of the invisible web into the discussion, but, as we can demonstrate, his calculation is misleading for academic text-based content.

Notes

1. http://scholar.google.com/
2. www.scirus.com
3. www.base-search.net/
4. www.vascoda.de/

References

Bergman, M.K. (2001), "The deep web: surfacing hidden value", *Journal of Electronic Publishing*, Vol. 7 No. 1, available at: www.press.umich.edu/jep/07-01/bergman.html (accessed April 6, 2006).

Brophy, J. and Bawden, D. (2005), "Is Google enough? Comparison of an internet search engine with academic library resources", *Aslib Proceedings*, Vol. 57 No. 6, pp. 498-512.

Jacsó, P. (2005), "Google Scholar: the pros and cons", *Online Information Review*, Vol. 29 No. 2, pp. 208-14.

Krause, J. (2003), "Standardization, heterogeneity and the quality of content analysis: a key conflict of digital libraries and its solution", paper presented at the IFLA 2003, World Library and Information Congress: 69th IFLA General Conference and Council, Berlin (International Federation of Library Associations and Institutions, Social Science Libraries Section), available at: www.ifla.org/IV/ifla69/papers/085e_trans-Krause.pdf

Lawrence, S. and Giles, C.L. (1999), "Accessibility of information on the web", *Nature*, Vol. 400 No. 8, pp. 107-9.

Lewandowski, D. (2005a), "Google Scholar – Aufbau und strategische Ausrichtung des Angebots sowie Auswirkung auf andere Angebote im Bereich der wissenschaftlichen Suchmaschinen", available at: www.durchdenken.de/lewandowski/doc/Expertise_ Google-Scholar.pdf (accessed December 13, 2005).

Lewandowski, D. (2005b), *Web Information Retrieval: Technologien zur Informationssuche im Internet*, DGI, Frankfurt am Main.

Lewandowski, D. (2005c), "Yahoo – Zweifel an den Angaben zur Indexgröße, Suche in mehreren Sprachen", *Password*, Vol. 20 No. 9, pp. 21-2.

Lewandowski, D. (2006), "Suchmaschinen als Konkurrenten der Bibliothekskataloge: Wie Bibliotheken ihre Angebote durch Suchmaschinentechnologie attraktiver und durch Öffnung für die allgemeinen Suchmaschinen populärer machen können", *Zeitschrift für Bibliothekswesen und Bibliographie*, Vol. 53 No. 2, pp. 71-8.

Lossau, N. (2004), "Search engine technology and digital libraries: libraries need to discover the academic internet", *D-Lib Magazine*, Vol. 10 No. 6, available at: www.dlib.org/dlib/june04/ lossau/06lossau.html (accessed April 6, 2006).

Lyman, P., Varian, H.R., Swearingen, K., Charles, P., Good, N., Jordan, L.L. and Pal, J. (2003), "How much information 2003?", available at: www.sims.berkeley.edu/research/projects/ how-much-info-2003/ (accessed April 6, 2006).

McKiernan, G. (2005), "Scirus: for scientific information only", *Library Hi Tech News*, Vol. 22 No. 3, pp. 18-25.

Mayr, P. and Walter, A.-K. (2005), "Google Scholar – Wie tief gräbt diese Suchmaschine?", paper presented at the 11. IuK-Jahrestagung: In die Zukunft publizieren: Herausforderungen an das Publizieren und die Informationsversorgung in den Wissenschaften, Bonn, May 9-11, available at: www.ib.hu-berlin.de/~mayr/arbeiten/Mayr_ Walter05-preprint.pdf (accessed April 6, 2006).

Notess, G.R. (2005), "Scholarly web searching: Google Scholar and Scirus", *Online*, Vol. 29 No. 4, pp. 39-41.

Ru, Y. and Horowitz, E. (2005), "Indexing the invisible web: a survey", *Online Information Review*, Vol. 29 No. 3, pp. 249-65.

Scirus (2004), "Scirus white paper: how Scirus works", available at: www.scirus.com/press/pdf/ WhitePaper_Scirus.pdf (accessed April 6, 2006).

Sherman, C. (2001), "Search for the invisible web", available at: www.guardian.co.uk/online/ story/0,3605,547140,00.html (accessed March 8, 2006).

Sherman, C. and Price, G. (2001), *The Invisible Web: Uncovering Information Sources Search Engines Can't See*, Information Today, Medford, NJ.

Stock, W.G. (2003), "Weltregionen des Internet: Digitale Informationen im WWW und via WWW", *Password*, Vol. 18 No. 2, pp. 26-8.

Williams, M.E. (2005), "The state of databases today: 2005", *Gale Directory of Databases*, Vol. 2, Gale Group, Detroit, MI, pp. XV-XXV.

Corresponding author
Dirk Lewandowski can be contacted at: dirk.lewandowski@uni-duesseldorf.de

WSIS, IFLA, UNESCO and GATS: networking for libraries on an international level

Christel Mahnke

Goethe-Institute, Munich, Germany

Abstract

Purpose – The World Summit on the Information Society (WSIS) shows the political dimension of the global information society. The purpose of this article is to consider the state of WSIS after the second phase and show possibilities for libraries to participate in the WSIS implementation. Furthermore, relations between WSIS, GATS and the "UNESCO convention on the protection and promotion of the diversity of cultural expressions" are explained. IFLA, as leading library organization, as well as national associations, is taking up the challenge of advocating libraries as agencies of information societies.

Design/methodology/approach – Provides a report and analysis of WSIS Summit 2003-2005 and WSIS Stocktaking for Libraries.

Findings – Relevance of WSIS Summit 2003-2005 and WSIS Stocktaking for Libraries.

Originality/value – The paper offers first-hand information on the WSIS process. It outlines proposals for networking between libraries and advocacy for libraries on global political level.

Keywords Information society, Libraries, Networking

Paper type Viewpoint

Networking and information politics: WSIS and other fields of action

For most librarians, networking is part of their daily life. On a national and international level, there are many joint projects and personal contacts made possible and kept alive by e-mail and the internet. However, libraries are only now becoming aware of their role in the global information society and the new field of information politics.

The World Summit on the Information Society (WSIS)[1] provides an excellent opportunity for librarians in the field of global politics. Since the 1990s, several world summits have been held by UN organizations to provide an opportunity to discuss, and hopefully solve, problem issues such as poverty and development, health and food. In 2001, the United Nations (UN) General Assembly decided to hold the first World Summit on the Information Society (WSIS). The first phase took place in Geneva in December 2003, and the second phase in Tunis, from 16 to 18 November 2005.

For librarians, it is easy to see why information should be of international political concern: the world is changing rapidly into a global information society. New technologies are shaping the way we live, learn, work and communicate. WSIS aimed to develop a clear statement of political will and to take concrete steps to establish the foundations for an information society for all, reflecting all the different interests at stake. For the first time, not only governments were part of the summit process, but all stakeholders: all relevant UN bodies and other international organizations, non-governmental organizations, the private sector, civil society, and the media.

Initially, WSIS was organized by the International Telecommunication Union (ITU), and focused on technical aspects, but with the United Nations Educational, Scientific and Cultural Organization's (UNESCO) active participation, WSIS no longer concentrated exclusively on technical matters, but rather on the cultural and educational impact of information. Also, the role of information in achieving the UN Millennium Goals was examined.

Parallel to the WSIS process, UNESCO actively pursued the development and endorsement of a "Convention on the protection and promotion of the diversity of cultural expression" (UNESCO, 2005). It was approved by the general conference of UNESCO in October 2005, and laid the foundation for national and international cultural policies.

Protection and development of cultural diversity is a challenge of our time. Societies are increasingly asserting their own cultural values, often firmly dissociating themselves from others in doing so. At the same time, however, mankind is confronted with its common global responsibility more than ever before. The coexistence of cultural fragmentation and globalization can cause tension, which in turn signifies a new potential for conflict. This calls for a worldwide cultural dialogue between partners with equal rights, the aim being to promote tolerance between various cultures and the appreciation of different national and regional cultural traditions. The Goethe-Institute, as one of the agencies of international cultural dialogue, took part in the shaping of the "UNESCO convention on the protection and promotion of the diversity of cultural expression".

General Agreement on Trade in Services (GATS) is a treaty that brings services into the irreversible process of opening world trade for merchandise and setting it free from regulations such as customs tariffs and other market entrance restrictions. All members of the World Trade Organization (WTO) are members of GATS. As far as telecommunication or energy markets are concerned, it makes sense to end state monopolies and allow international competition into domestic markets. However, GATS can be seen as running against national and international cultural policies. In several ways, it can counteract political decisions like providing free schooling for children or inclusive library services. Therefore, librarians and other members of civil society in Germany, such as media or education workers, have spoken in favor of a clear distinction between commercial and public spheres.

The European Bureau of Library, Information and Documentation Associations (EBLIDA), has issued a statement on GATS. According to EBLIDA, libraries or educational institutions risk losing their support from public funds in the country in which they operate, if service markets are open to international competition (EBLIDA, 2005).

WSIS and libraries: Geneva 2003

The International Federation of Library Associations and Institutions (IFLA) took part in the WSIS process from the very beginning (Haavisto and Lor, 2006). As the leading international body representing the interests of library and information services and their users, IFLA representatives managed to put libraries and their contributions to the global information society on the summit's agenda. The Goethe-Institute, Germany's cultural institution abroad, also took part in the summit, as a part of German and international civil society and as member of IFLA.

In November 2003, just before the first phase of WSIS took place in Geneva, IFLA held a preparatory conference under the title "Libraries @ the Heart of the Information Society" (IFLA and SLIR, 2003). Thanks to intense lobbying in the crucial days before the conference, and perfect preparation by Swiss librarians in Geneva, libraries are mentioned in the Geneva Declaration of Principles and the Geneva Plan of Action as important partners for the realization of a democratic and egalitarian information society.

The range of recommended actions for governments covers the connection of rural libraries to information and communication technology (ICT) networks, the creation and development of digital public library and archive services, reviewing national library strategies and legislation, developing a global understanding of the need for "hybrid libraries", and fostering worldwide cooperation between libraries. Promotion of e-literacy skills for all is seen as a future task for all governments, taking advantage of existing facilities such as libraries.

Specific training programs in the use of ICTs in order to meet the educational needs of information professionals, such as archivists, librarians and other relevant professional groups are also mentioned. Training of information professionals should focus not only on new methods and techniques for the development and provision of information and communication services, but also on relevant management skills to ensure the best use of technologies. Development of content is seen as equally important. All in all, the Geneva plan of action is a perfect basis for libraries to develop national and regional strategies and projects.

Unfortunately, not all governments are inclined to follow the plan they agreed upon. In Germany, there is no national library strategy, and the place of libraries is not "@ the heart of the information society", at least not in the eyes of politicians. So far, they are hardly mentioned in the European Union programs for ICT development. But the libraries themselves have taken up the challenge to show their contribution to the implementation of the WSIS goals. In 2004, IFLA set up a Success Story Database[2]: a showcase of libraries as access points to information, ICT learning centers and agencies for continuing education. The database is searchable online and available on CD-ROM. Many projects are cooperative activities of three or more partners, quite often from different countries.

On the way to Tunis

Librarians watched closely the preparatory process of the second phase, namely a series of preparatory conferences and meetings of WSIS working groups. These groups were set up to find solutions for the most important remaining problems of the first phase: internet governance and financing of Internet connections in developing countries. But not only governments were preparing for the second phase. Civil society was equally active, discussing, in Germany, topics such as open access, the importance of information commons, bridging the digital divide and gender mainstreaming in the field of ICT. The Heinrich Böll Foundation, a legally independent political foundation affiliated with the Green Party, provided a platform for German Civil Society[3]. A prominent civil society figure on the international stage has been the Communications Rights in the Information Society (CRIS) campaign[4]. CRIS has been furthering civil society participation at the WSIS. Other important players, to name but a few examples from a multitude of groups and organizations, have been the OpenWSIS Initiative and

the Conference of Non-Governmental Organizations in Consultative Relationship with
the United Nations (CONGO).

In November 2005, IFLA held the pre-Summit Conference "Libraries – the
information society in action"[5], in Alexandria. The choice of Bibliotheca Alexandrina
as the location for IFLA's conference has a special significance: it was the first major
IFLA event ever to take place in the Arab world. Contacts between IFLA and the Arab
Federation for Libraries and Information (AFLI) were renewed and will hopefully lead
to more co-operation with libraries and other civil society members in the Arab world.
The attention and respect librarians gained from official WSIS organizers was reflected
in the fact that IFLA's pre-Summit Conference was one of three official side events of
the second phase of the World Summit on the Information Society.

The conference was intended to show that libraries and information services
provide societies with a ready-made tool to further the information society and achieve
the UN Millennium Development Goals. At the end of the conference, the "Alexandria
proclamation on information literacy and lifelong learning" (IFLA, 2005a) and the
"Alexandria manifesto on libraries – the information society in action" (IFLA, 2005b)
were launched. The manifesto calls on governments and non-governmental
organizations (NGO) to invest in library and information services as vital elements
of the information society strategies, to support unrestricted access to information and
freedom of expression, to promote open access to information and to recognize the
importance of information literacy. Obviously, this call did not go unheard: IFLA
president Alex Byrne was invited to speak in the plenary session of the WSIS closing
session, as one of just a few representatives of civil society.

Tunis: information society with restricted access

It is a known fact, that many states in the Arab world, including Maghreb states, have
restricted freedom of information[6]. Internet-access is filtered, and some sites by
human rights groups are not accessible at all. People are prosecuted if they look up
"forbidden" sites. However, Tunisia, as the WSIS host country, welcomed its guests as
"the land of civilization, culture and enlightened thinking"[7]. Unfortunately, not all
guests were welcome. Some journalists were not allowed to enter the country, others
were restricted in their work (IFLA/FAIFE, 2005a). The summit itself, which counted
about 20,000 participants, was held on a site far from town and under strict isolation
from local citizens. Security concerns were the official reason, but it was unclear whose
security was at stake. These events did not come as a surprise to librarians, who were
aware of restricted freedom of information thanks to the recent report of the IFLA
Committee on Free Access to Information and Freedom of Expression (IFLA/FAIFE)
on Tunisia (IFLA/FAIFE, 2005b).

Some civil society side events, one of which was to take place in the
Goethe-Institut, had to be cancelled because Tunisian police would not allow
participants to enter. These events only underlined the importance of freedom of
information. All European delegations expressed their solidarity.

There were many librarians present at the conference, some as official delegation
members of their governments. They seized on the rich networking possibilities and
promoted libraries' achievements and the services they provide in making knowledge
and information available.

Summit conferences are much more than plenary sessions and working groups. The exposition Information and Communication Technologies for Development (ICT4D) was a fascinating showcase of best practice and new ideas from all parts of the world: the 100-dollar-laptop, developed for poor countries by the Massachusetts Institute of Technology (MIT), was on display, as well as interactive educational websites and mobile health care units for remote areas with internet access to the best hospitals of the capital. Many development agencies and nearly all UN organizations were present, all in all 250 exhibitors and more than 350 conferences. IFLA had its meeting point at the Global Knowledge Partnership (GKP)[8] stand, the leading international multi-stakeholder network committed to harnessing the potential of information and communication technologies for sustainable and equitable development. The members of this widespread network, ranging from grassroots practitioners to policy-makers, gave interesting insights into their work for development.

Open questions after Tunis

The three main themes of this summit were Internet governance, financing of the future development of the internet and, of course, the follow-up to the summit. As to internet governance, no real solution could be found to the question of who should control the Internet. For the time being, power stays with the Internet Corporation for Assigned Names and Numbers (ICANN), a private company based in the US. A new body for multi-stakeholder policy dialogue called the Internet Governance Forum (IGF) was set up after the summit to have a closer look at ICANN activities and decisions. Its first meeting will take place in October 2006.

Finance matters were also a point of disagreement. Poor countries called for a digital solidarity fund, but donor countries were rather inclined to use existing development aid funds for the development of internet infrastructure and capacity building in the field of ICT. Looking back to Geneva, there was at least a growing awareness that commercial interest alone will not be enough to secure convenient and affordable internet access in all parts of the world. This is best seen in ITU's slogan for Tunis:

We always thought communication a human need – now we know it is a human right.

Civil society members parted from Tunis with mixed feelings. Some judged the results as too superficial, others pointed out the significant presence and lively contribution of civil society. However, the integration of all stakeholders in the WSIS process is seen as a success and will continue in the implementation process.

There are different instruments for WSIS stocktaking, such as a database of all WSIS-related projects (2,648 projects as at 10 March), several websites and a golden book intended to promote new commitments and initiatives that stakeholders have announced, during the Tunis phase of the WSIS.

Next steps for librarians

The summit has taken place, but the process of implementation is just beginning. Librarians should ensure that they capitalize on the formal recognition, general awareness, and advocacy experience that were gained. In Europe, new initiatives like i2010 – A European Information Society for Growth and Employment[9] and i2010 Digital Libraries Initiative[10] are being launched. During the public consultation,

librarians gave their opinions on this new information society strategy. In March 2006, a high level expert group on digital libraries was established. The group will advise the European Commission on how to tackle key challenges in making Europe's cultural heritage available online.

Over the next few years, we will have to develop new ideas and tools for library advocacy at every level of society. Networking on a European and a global scale will be crucial if we are to make an impact. It is more necessary than ever, that librarians speak with one voice. The German Federal Union of German Library and Information Associations (Bibliothek and Information Deutschland – BID) brings together libraries, librarians and information experts with supporting institutions. One of BID's main fields of action is international co-operation in the field of librarianship.

IFLA will remain the most important partner on an international level. German participation in IFLA activities is lively and will surely increase over the coming years, with Claudia Lux as the incoming president. UNESCO, with its impressive range of library projects, is also a good and reliable partner for the shaping of tomorrow's information societies. The UNESCO libraries portal[11] gives an overview and project information.

Librarians are active in building a user-friendly, inclusive information society. Through their networks, they are linked to all layers of society and can influence information policy on a local and a global level.

Notes

1. www.itu.int/wsis/
2. http://fmp-web.unil.ch/IFLA/
3. www.worldsummit2003.org/
4. www.crisinfo.org/
5. www.bibalex.org/wsisalex/agenda.htm
6. www.grc.ae/?sec=Arab+Human+Development+Report&sec_type=h&sub_opt=181
7. www.smsitunis2005.org/plateforme/index.php?lang=en
8. www.globalknowledge.org/
9. http://europa.eu.int/information_society/eeurope/i2010/index_en.htm
10. http://europa.eu.int/information_society/activities/digital_libraries/index_en.htm
11. http://portal.unesco.org/ci/en/ev.php-URL_ID=6513&URL_DO=DO_TOPIC&URL_SECTION=201.html

References

European Bureau of Library, Information and Documentation Associations (EBLIDA) (2005), "Statement on the proposal of the European parliament and of the council for a directive on services in the internal market", November 2005, available at: www.eblida.org/position/ServicesDirective_Statement_Nov05.htm (accessed April 5, 2006).

Haavisto, T. and Lor, P.J. (2006), "IFLA in action at WSIS", available at: www.ifla.org/III/wsis/WSIS-report27-01-2006.html (accessed April 5, 2006).

IFLA Committee on Free Access to Information and Freedom of Expression (IFLA/FAIFE) (2005a), *IFLA Protests Crack Down of Intellectual Freedom in Tunisia*, press release, November 18, available at: www.ifla.org/V/press/Pr-18112005.htm (accessed April 5, 2006).

IFLA Committee on Free Access to Information and Freedom of Expression (IFLA/FAIFE) (2005b), "Report on IFEX – TMG mission to Tunis", available at: www.ifla.org/faife/faife/tunis-report2005.htm (accessed April 5, 2006).

IFLA and SLIR (2003), "Libraries @ the heart of the information society", *Proceedings of the IFLA Pre-World Summit Conference*, Geneva, November 3-4, available at: www.ifla.org/III/wsis/proceedings2003.pdf (accessed April 5, 2006).

International Federation of Library Associations and Institutions (IFLA) (2005a), "Alexandria proclamation on information literacy and lifelong learning", available at: www.ifla.org/III/wsis/BeaconInfSoc.html (accessed April 5, 2006).

International Federation of Library Associations and Institutions (IFLA) (2005b), "Alexandria manifesto on libraries, the information society in action", available at: www.ifla.org/III/wsis/AlexandriaManifesto.html (accessed April 5, 2006).

United Nations Educational, Scientific and Cultural Organization (UNESCO) (2005), "Convention on the protection and promotion of the diversity of cultural expression", October, available at: http://unesdoc.unesco.org/images/0014/001429/142919e.pdf (accessed April 27, 2006).

Corresponding author
Christel Mahnke can be contacted at: mahnke@goethe.de

Impact measures for libraries and information services

Roswitha Poll

University of Münster, Münster, Germany, and

Philip Payne

Birkbeck College, University of London, London, UK

Abstract

Purpose – The purpose of this article is to demonstrate the importance of impact/outcome research in libraries.

Design/methodology/approach – The paper gives an overview of purposes and methods used in impact research and illustrates this through project experiences.

Findings – Various projects worldwide are trying to prove that use of library services can positively influence skills and competences, attitudes and behaviour of users. The benefits that users experience by using library services can be assessed in terms of knowledge gained, higher information literacy, higher academic or professional success, social inclusion, and increase in individual well-being.

Research limitations/implications – The main problem of impact research is, that influences on an individual are manifold and that therefore it is difficult to trace changes and improvements back to the library. The paper shows methods that are tested and used at the present. More investigation is needed to identify methods that could be used to show a library's overall impact or to develop measures that would permit benchmarking between institutions.

Practical implications – The paper shows practical examples of impact assessment, covering "soft" methods like surveys, interviews, focus groups, observation and quantitative methods like tests, analysis of publications, or usage data.

Originality/value – The paper acquaints libraries with a topic that is not yet well known and, by showing practical examples, demonstrates how libraries can attempt to assess their impact.

Keywords Libraries, Information services, Research

Paper type Research paper

The search for an ultimate measure of benefit may be illusory (Revill, 1990).

The need for evidence

The value of libraries for the individual and for society has long been seen as self-evident. However, in times when users are becoming increasingly independent in their information seeking, when information seems to be free on the web (even where libraries have paid for access), and physical visits to libraries may decrease, the benefits gained from funding libraries are questioned not only by funding institutions but also by the public. Funding libraries is an expensive business: print and electronic collections, buildings and equipment, and especially staff costs, constitute a considerable factor in the budget of universities or communities. The questions are:

(1) Does investment in libraries represent value for money?

(2) Are there tangible, demonstrable effects arising from library use?

(3) Do such effects serve the goals of the funding institution?

(4) Could such effects be achieved without the existence of the particular library?

In the competition for scarce resources, it becomes vital for libraries to show evidence of the impact and value of their services, preferably in quantified results. "Outcome-based evaluations can be thought of as an accountability measure ..." (Hess and Klekotka, 2005, p. 272). Librarians themselves are of course convinced of library benefits. In the Alexandria Manifesto, adopted in 2005 in preparation for the World Summit on the Information Society, the International Federation of Library Associations and Institutions (IFLA) sees library benefits in the following fields (IFLA, 2005):

· democracy;

· intellectual freedom;

· information literacy;

· information equality;

· reduction of poverty; and

· cultural diversity.

But how can we prove what we believe?

Impact of libraries

Libraries have always been able to calculate the inputs into services (funding, staff, collections, space, equipment) and have become increasingly sophisticated in measuring the outputs of those services (e.g. loans, visits, downloads, reference transactions). Measures have also been developed to assess the quality of library services and the cost-efficiency of the library's performance. But quantity of use and quality of performance do not yet prove that users benefited from their interaction with a library. Measuring impact or outcome means going a step further and trying to assess the effect of services on users (Figure 1).

"Impact" and "outcome" are often used synonymously in the professional literature. "Value" or "benefit" are generally broader terms. Definitions of library outcome generally highlight the effect on individual users or on users collectively. Impact links this with the library's aims and objectives and their relationship with its host institution's goals. Revill (1990) describes outcomes as follows:

Outcomes can be seen as the eventual result of using library services, the influence the use had, and its significance to the user (p. 360).

This is echoed by the Association of College and Research Libraries in the USA:

Outcomes are the ways in which library users are changed as a result of their contact with the library's resources and programs (ACRL, 1998).

The existence of a library and the use of its services can effect changes in skills, competences, attitudes and behaviour of its users, similar to changes effected by other cultural institutions like museums or archives.

Generally speaking, outcomes of cultural institutions include:
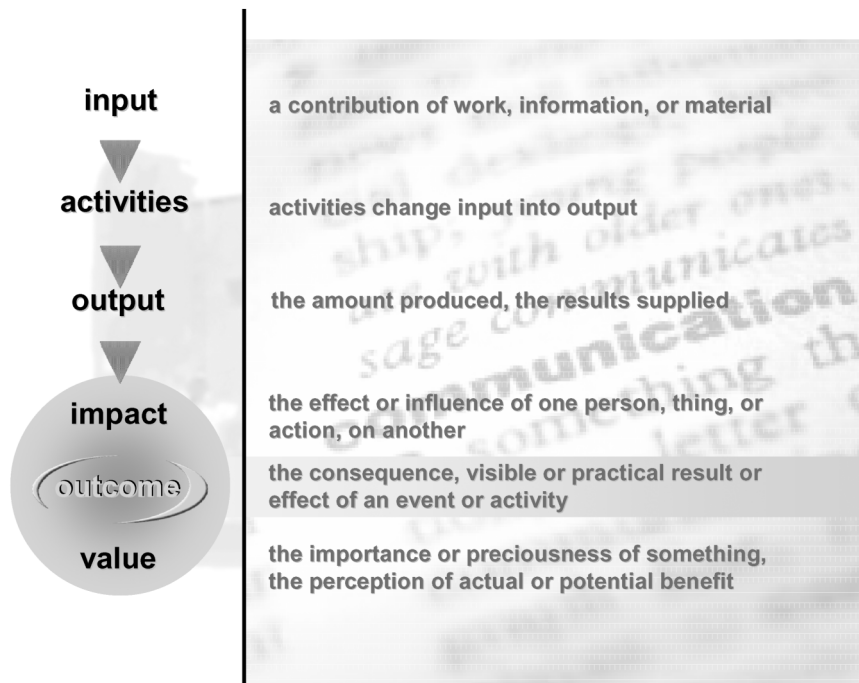
· knowledge;

· information literacy;

Figure 1.
From input to
impact/outcome

- higher academic or professional success;
- social inclusion (e.g. of elderly people or immigrants); and
- individual well-being.

Such outcomes may appear in different ways. Hopefully, they will be positive, but there is also the possibility of negative outcomes: a bad experience during a library visit can result in a negative attitude to library use. Outcomes may be direct, immediate (finding useful information) or long-term (gaining information seeking competences). Actual benefits to the individual user differ from the potential benefits, the preciousness of a library's existence for issues like local culture or children's literacy. Finally, there may be intended and unintended outcomes. Intended outcomes are those that the library aims at according to its mission and goals. Unexpected outcomes – like people finding social contacts in the library – can nevertheless further the positive attitude to library use.

Outcomes, as research shows:

- are not always predictable;
- are generally rather an addition to previous experience than a radical change in attitudes;
- will be higher if a gain in skills and competences or a change in behaviour seems promising to the user; and
- often become visible only in long-term development.

Impact planning
Most libraries are part of an institution – a community, a school, a commercial firm, or a university – and the library's mission and goals must be adjusted to those of its parent institution. The impact of library services should therefore support the institutional goals and the strategic aims of the institution.

A good example is the impact of libraries compared to the goals of universities. What universities want to achieve can be summarized thus:

(1) Recruitment and retention of students and excellent academic staff.

(2) Effective teaching, resulting in:
   · high graduation rates;
   · high grades in examinations; and
   · high employment rates after examinations.

(3) Effective research, resulting in:
   · high valuation and use of research results and publications;
   · high amount of special grants; and
   · status, awards, honours.

In summary, universities aim at achieving independence of thinking and judgement, competent use of information, thorough professional knowledge for their students, and research results of high relevance. Most of these goals can be supported by library services, and libraries should try to prove the connection between use of their services and the institution's success.

Assessing impact – the problems
Trying to show an impact of libraries on individuals and society is obviously a much more difficult venture than counting outputs. The following problems appeared in most studies:

· A service can have different value and outcome for different user groups. A training session in special databases will have less effect on freshmen than on postgraduates who need these resources directly for their work.

· Data that could be relevant for demonstrating impact are not available because of data protection rules (e.g. individual data about grades in exams).

· The data or correlations found in projects until now are in most cases not comparable, as differing methods were used. Standardization of methods will be necessary to allow for benchmarking of results.

· Long-term effects can often not be assessed if the users are no more available for tests or surveys.

· All methods that have been tested until now are time-consuming.

But the most challenging problem is that it is nearly impossible to separate library impact from other influences and to prove that changes in competences or behaviour are indeed an effect of using library services. Influences on individuals are manifold. Users may have gained information and competences from friends or teachers, from using media outside the library or searching the Internet. As it is often not possible to

find positive proof of a direct influence of the library, surrogate measures must be used that at least indicate some influence. These problems are most troublesome when assessing the overall impact of a library and its services. They become less apparent in evaluating the outcome of one single activity like implementing a new service or conducting a user training programme, as the behaviour and skills before and after the implementation or the training can be more readily measured. Another difficulty is that it may take time before the impact of a library's interventions becomes clear. As Everest and Payne (2001) state:

> Assessing impact is not easy and it is not an exact science. We are dealing with a changing environment where people, services, and needs are constantly evolving. Any research will inevitably provide a snapshot of what is happening at a particular point in time (p. 21).

Methods for assessing impact
The ways that have been used for showing impact can be differentiated into quantitative and qualitative methods. Quantitative methods try to measure changes in competences or behaviour or to find correlations between library use and a person's academic or professional success. The following methods have been used:

- tests that assess user skills before and after a training or the use of a library service;
- performance monitoring/data mining;
- unobtrusive observation;
- analysis of citations in course work or research publications; and
- comparison of success data with use data.

Qualitative ("soft") measures, developed in social science, try to assess outcomes by evaluating users' experiences and opinions. Methods used are:

- surveys (print, telephone, or online);
- interviews;
- focus groups, discussion groups; and
- users' self-assessment of skills and competences gained.

Exit surveys (surveys conducted on leaving a service) can be especially useful as they can immediately catch users' impressions about having benefited. Self-assessment has proved less reliable, as users tend to rate their own competences somewhat higher than they really are. Qualitative methods supply a rich fund of "stories" about personal experiences and judgements. These stories must be organized in order to recognize patterns and crucial points and – if possible – to show statistics of outcomes as to user groups. The results of qualitative methods will of course have a subjective bias; they show the "perceived outcome". They should therefore be compared with results of quantitative methods or with statistics of library use in order to validate the results. But the "anecdotal evidence" will be invaluable in reporting to the public and the institution, as it serves to make statistics understandable and believable.

There is a debate in outcome research as to whether user satisfaction could serve as an outcome measure. As user satisfaction surveys are now well established in libraries,

it would be easy to use the results for assessing the impact of libraries. But is satisfaction with the library or with a library service indeed an outcome in the sense of benefits? Opinions are divided:

> Satisfaction on the part of a user is an outcome. So is dissatisfaction. The task force considers simple satisfaction a facile outcome, however, too often unrelated to more substantial outcomes that hew more closely to the missions of libraries and the institutions they serve (ACRL, 1998).

> Customer satisfaction [...] is neither outcome nor output. Rather, it is a qualitative assessment of library outputs [...] (Cram, 2000, p. 23).

User satisfaction surveys ask for a user's direct or long-term experience with library services or for experience compared to expectation. High satisfaction could mean that the library has been effective in conveying the view: it is well worth to use a library. But this does not mean that there is already a change in skills, competences, and behaviour. User satisfaction could rather be seen as giving a good basis for such changes in furthering receptivity and thus rendering outcomes possible.

The topics of impact assessment
Projects of impact assessment can be grouped as to the following topics:

- correlation of library use and academic or professional success;
- the library's impact on information literacy;
- the importance of the local library for research;
- the social impact of libraries; and
- the financial value of libraries.

Correlation of library use and academic or professional success. Statistics of the frequency and intensity of library use are compared to data of individual success. Research in this area has concentrated on academic success of students. Library use is expressed in data about:

- the frequency of use (e.g. of loans or visits);
- the range of services used (e.g. reference service, user training, interlibrary loan);
- the range of collections used (e.g. not only the undergraduate collection but e-journals, special material); and
- attendances at training sessions.

Academic success is expressed in terms of:

- short studying times;
- high grades in examinations;
- student retention;
- high employment rate after examinations; and
- the quality of publications (high impact factor, publishing in peer-review journals).

Projects have not always found a statistically valid correlation between library use and academic success. The most interesting results are probably those of the Department of Information and Library Studies of the University of Cape Town (De Jager, 1997, 2002). The projects tried to find out whether students had only used the basic literature in "short loan material" or had gone on to the more special literature in the general collection. When comparing the data not of individuals but of groups, the results showed that students with a high grade in examinations had indeed used the specialized literature to a higher degree. But such data related to individuals will not be available everywhere because of data protection rights. And the question remains whether the quick and successful examination was (also) an effect of library use.

The library's impact on information literacy. Outcome research in libraries has concentrated very much on this issue, due to the importance of showing "learning outcomes" in education. Libraries are more and more integrated in information literacy teaching in universities, especially in bachelor/master studies. The new role of the "teaching library" calls for assessment of teaching results. Methods for measuring outcomes of user training have been already well tested and even standardized. They include:

- surveys after instruction;
- pretest/post test;
- self-assessment of users;
- behavioral observation;
- transaction logs;
- longitudinal studies; and
- assessing changes in students' bibliographies.

Tests, especially combinations of pretest and posttest, are the preferred method (Brown and Krumholz, 2002; Dunn, 2002; Fister, 2003; Mark, 2004). In the last years, citation analysis has also been used for assessing changes in users' information literacy competences (Beile *et al.*, 2004; Middleton, 2005; Tuñon and Brydges, 2005). Whatever method was used, it seemed in most cases possible to show direct impact of library training and services on information literacy.

The importance of the local library for research. Citation analysis is also used for showing a library's value for research in its institution (Ahtola, 2002; Smith, 2003). The citations in bibliographies of dissertations, research papers or students' papers are compared with the library's print and electronic collection. The question is: what percentage of the material cited was (or could have been) retrieved via the local library?

In order to check whether the materials were indeed taken from the library collection, a survey to, or interview with, the authors might be added. Questions could be:

- Where did you find the material you cited in your publication?
- Could you have achieved a similar result without your local library?

The University and Regional Library Münster, Germany, examined 7016 citations in 20 doctoral dissertations and found that 70.8 percent would have been available via the library. A survey to the authors corroborated the data: 15 of the 20 authors said they

had frequently retrieved their information from the local library system (Kayß and Poll, 2006). This method is certainly one of the "surrogate measures" for assessing library outcomes.

The social impact of libraries. In this complex of questions, libraries try to show the value of libraries for the individual (direct benefit) or society (indirect benefit). Methods used are interviews, surveys or focus groups with actual and potential users. In most cases questions as to sociodemographic data (age, gender, ethnic group, income, employment status, educational grade) and as to library use (frequency, services used) are added in order to have a background for recognizing patterns in the answers.

Users and non-users are asked for:

- direct benefit from a library use;
- potential benefit for the interviewee by this library;
- indirect (potential) benefit of a library's existence (e.g. free access to information, cultural life in the community, children's literacy, social inclusion); and
- potential value of the library for future users.

Projects for assessing social impact have been frequent in public libraries (Bohme and Spiller, 1999; Debono, 2002; Linley and Usherwood, 1998).

Results of social impact projects should be compared with statistics of library use and library user groups that can show developments in the choice of services used and the percentages of active users out of certain groups (e.g. fringe groups).

The financial value of libraries. Measuring the value of libraries in terms of money is certainly the issue of impact assessment that would be most interesting to the funding institutions.

As most library services have no equivalent on the common market and therefore no "market prices" can be determined, two other ways have been tried for assessing an economic value:

- assessing time costs ("replacement value of a client's time"); and
- using the contingent valuation method.

The time costs method is based on the assumption: Users invest time and effort in order to use library services. The value that they – or their institution – place on that use must be at least as high as their "sacrifice" of time. Time costs are calculated by multiplying users' time spent with library services with the average salary costs of the population served by that library. The method has been used in special libraries, but does not fit, e.g. student populations.

The contingent valuation method has been developed in order to assess the financial value of non-profit organizations and services, especially projects in health care, environmental protection, education or culture. Persons directly or potentially interested in such services are asked to rate the value in financial terms, expressed by their:

- Willingness-to-pay: what would you pay for maintaining this library/this special library service?
- Willingness-to-accept: which sum would you accept as an equivalent if this library/this special library service were given up?

In the surveys or interviews, people are given options between sums they would pay (e.g. in higher taxes) or accept (e.g. in lower taxes). The difficulty in such surveys is that people are asked to financially rate services or institutions that they never thought of in terms of money. In the answers, the willingness-to-pay is often somewhat lower than the willingness-to-accept, as people are afraid to name a sum they would pay. Examples for the use of the contingent valuation method are already frequent in the sector of public libraries (Morris *et al.*, 2001; Holt *et al.*, 1999; Aabø, 2005). The best-known example is that of the British Library (Pung *et al.*, 2004) that came out with the result: For every £1 of public funding the British Library receives each year, £4.4 is generated for the economy. Such data are indeed invaluable in reporting.

Using the results
Results of impact assessment can be used for several purposes:

- accountability, justification of resources used and services offered;
- decision making and resource management;
- improvement of services in order to achieve higher outcomes; and
- benchmarking with libraries of similar structure and mission.

But the most urgent issue is to promote the library's role, to show what one library, what all libraries can do for their users and society. Libraries are too often forgotten in legislation, in community or institutional planning, or when setting priorities in funding. They should actively promote the benefits derived from their services and substantiate such statements with the evidence of data and lively stories. "Libraries on the agenda" will be the presidential theme of the next IFLA president, Claudia Lux; impact assessment can help to put libraries in the limelight.

Assessing the value of libraries is no easy task. Intangible issues like the impact on knowledge, competences and behaviour of persons must be made visible and understandable. "Searching for the most elusive indicators" (De Jager, 2001), "Six impossible things before breakfast" (Cram, 2000): papers on outcome measures show the difficulty.

The LIRG/SCONUL Impact Initiative: practical examples of assessing impact
Background to the initiative
The LIRG/SCONUL Impact Initiative has been concerned with assessing the impact of higher education libraries in the UK on learning, teaching, and research. In the context of the initiative, "impact" has been seen in terms of whether libraries have made a difference as a result of a variety of different interventions. The initiative has then sought to develop methodologies that can be used by institutions to measure the difference that they have made.

The initiative has not sought to assess a library's overall impact. It has not been looking, for example, at a library's impact on student achievement. Instead, a project-based approach has been adopted. Individual institutions participating in the initiative have been investigating their impact in relation to particular services or innovations. The areas that participating institutions have investigated usually relate to the library's strategic goals and to their change agendas. Not surprisingly, the choice

of topics to investigate has often been guided by the strategic aims of the host institution. The libraries have sought to show their contribution to the work of their institutions. The initiative has been concerned with developing methodologies that can be used to assess a library's impact.

The initiative overall has effectively been an umbrella for a number of institution-based projects that have been conducted at the same time using a common approach. By taking a co-ordinated approach, we have sought to encourage the sharing of experience between those institutions participating in the Initiative. Participants in the initiative have been supported by facilitators. The facilitators have been consultants with expertise in assessing library impact (Markless and Streatfield, 2005). They have assisted in setting up the institutional impact projects using an "impact process model" that they had developed in previous work, and reviewing progress. Most of the support has been delivered through workshops that have been run at the start, midway, and end of the programme.

When the initiative began, it had been hoped to develop sector-wide impact measures but this was found to be over-optimistic. It soon became clear that, in view of the different institutional contexts and priorities, this could not be achieved within the timescales of the initiative. When more experience has been built up of assessing impact, benchmarking impact between comparable institutions might be possible.

Participants in the initiative
Who has been involved with the initiative? The Library and Information Research Group (LIRG) is a special interest group of the UK's Chartered Institute of Library and Information Professionals (CILIP). LIRG aims to promote the value of information research and to link research with practice. The Society of College, National and University Libraries (SCONUL) has provided support through its Working Group on Performance Improvement. As indicated earlier, the initiative has been facilitated by David Streatfield and Sharon Markless of Information Management Associates (IMA). IMA have a strong track record in assessing library impact, primarily in relation to health libraries, further education college libraries, and school libraries. They also have considerable research experience and expertise in research methodologies.

The other participants in the Initiative have been the institutions themselves. There have been two phases of the initiative. Phase 1 ran from July 2003 until July 2004 and involved ten institutions (Table I). Phase 2 ran from July 2004 until December 2005 and involved 12 institutions (Table II). All participating institutions responded to a call for volunteers. We were seeking a good range of institutions – institutions in different parts of the UK, of different sizes, and of different types.

The two sponsoring organisations, LIRG and SCONUL, contributed a small amount of pump-priming funding at the outset. Otherwise, the costs of participation have been covered by the libraries that have volunteered to be involved in the initiative. Pooling resources and benefiting through combined efforts was therefore a major feature of the Initiative.

Process adopted by the initiative
Participating institutions were required to identify an area where they wished to assess their impact. Although we were interested in covering a wide range of topics, a convergence emerged early on which led to eight of the ten institutional projects being

Birkbeck College	Impact of an online induction tutorial
University College Chester	Impact of investment in electronic resources
Glasgow Caledonian University	Information literacy strategy: awareness of the strategy and its impact
University of Gloucestershire	Impact of an online information skills tutorial
University of Leeds	Impact of information literacy initiatives
Leeds Metropolitan University	Information literacy strategy: awareness of the strategy and its impact
Northumbria University	Impact on improving students' confidence and competence in information and IT skills
Open University	Information literacy programme: impact on information literacy amongst students
University of Teesside	Impact of library support to partner institutions
University of Warwick	Impact of the library on the research process

Table I.
Institutions in phase 1 of the LIRG/SCONUL impact initiative

University of Birmingham	Institutional repository and scholarly communications advocacy
Bournemouth University	Equality of access to e-resources
Brunel University	Information skills for research postgraduates
University of Edinburgh	Subject-specific electronic resources and the research process
University of Glasgow	Subject-specific electronic resources and the research process
King's College	Integrated enquiry desks
Newman College	Support for widening participation
University College Northampton	Support for widening participation
University of Southampton	Library support for non-traditional students of modern languages
Staffordshire University	Single point of access for IT and library enquiries
University of Stirling	Student use of subject resources web pages
University of the West of England	Increasing amounts spent on electronic information services (EIS) each year

Table II.
Institutions in phase 2 of the LIRG/SCONUL impact initiative

concerned with information literacy. Other topics covered in phase 1 were e-resources, impact of library support on the research process, and the impact of library services to partner institutions. As the focus of phase 1 had been so strongly on information literacy, we were determined to cover a wider range of topics in phase 2. Five of the twelve institutions looked at e-resources, two looked at support for institutional widening participation initiatives, and two looked at the impact of converged library and IT help desks. Other projects covered library support for non-traditional students of modern languages, information skills for research postgraduates, and advocacy for institutional repository and scholarly communication. A special issue of LIRG's *Library and Information Research* provides overviews of phase 1 (Payne and Conyers, 2005; Markless and Streatfield, 2005; Blagden, 2005) plus reports from each of the ten participating libraries.

All of the projects followed a common approach that involved:

· choosing an area where the library was seeking to assess impact;

· articulating objectives that set out what the library was seeking to achieve in the chosen area;

· developing success criteria against which a judgement can be made as to whether the objectives have been met;

- creating impact measures for the chosen area of investigation;
- identifying evidence that needed to be collected in respect of these impact measures; and
- selecting appropriate research methods to collect that evidence.

The following example illustrates the stages that were followed. Learning Support Services at Leeds Metropolitan University had developed an information literacy framework and wanted to assess the impact of the framework on learning and teaching in the university. For each objective of the information literacy framework, they identified success criteria by which they could judge whether or not the library had been successful in what it had set out to do. They then chose the research methods that they would use in order to gather the evidence of whether or not they had been successful (Everest *et al.*, 2005). These objectives in respect of an information literacy framework were shared with Glasgow Caledonian University (Crawford, 2006).

Information Literacy Project (Leeds Metropolitan University)

(1) *Objective 1.* To raise awareness of the importance of information literacy amongst students and academic staff.

- Success criteria. Every member of academic staff aware of the information literacy framework.
- Evidence and collection method. Learning advisers to carry out personal or telephone interviews to assess level of awareness.

(2) *Objective 2.* To equip students with the skills to become independent seekers and discerning users of information in their studies.

- Success criteria. More diverse range of information sources cited in bibliographies.
- Evidence and collection method. A team of learning advisers to review student bibliographies before and afterwards in partnership with selected academic staff.

(3) *Objective 3.* To integrate information literacy into the student curriculum.

- Success criteria. Evidence that information literacy is increasingly integrated within the curriculum.
- Evidence and collection method. To use an information literacy teaching audit to assess current level of integration.

Methods used to collect evidence

What sorts of methods were used by participants in the initiative? Statistics on usage were important for many of the projects. Often these already existed and did not require additional data collection. There was particular interest in using statistics on the use of e-resources. Questionnaires, interviews, and testing students' skills levels after library interventions were frequently used. However, the projects demonstrated considerable diversity in the data collection methods used. The methods chosen tended to have an emphasis on qualitative "soft" methods. The emphasis on softer methods was not surprising as the initiative was based on action research. Markless and Streatfield (2006) highlight the appropriateness of this approach to the initiative:

Action research as a form of social research is not a detached specialised, technical activity but one closely linked to reflective practice, designed to be undertaken by practitioners and empower them (p. 4).

Outcomes of the initiative
What did the participants achieve through participation? Institutions reported very different experiences of participation in the initiative. Some struggled and others found it deeply rewarding. However, many of the projects reported that the initiative led to:

- deeper understanding of how the library supports academic processes;
- improved dialogue with academics and stakeholders;
- misconceptions (amongst library staff and amongst stakeholders) being challenged;
- the development of library staff – particularly in respect of their understanding of the contribution of the Library to learning, teaching, and research – but also in developing new skills; and
- a raised profile for the library within the organisation.

The outcomes for individual institutions have been interesting. Everest *et al.* (2005), for example, describe how the information literacy framework at Leeds Metropolitan University was produced as a booklet and circulated to all academic staff. Learning advisers subsequently conducted telephone interviews and found that over 50 per cent of academics had received the booklet and read it. However, around 20 per cent of those who had received the booklet had not read it. The project team also analysed bibliographies from student dissertations, developing criteria for what constitutes a "good" bibliography with academic staff. They report that initial results show that information skills teaching had lead to an improvement in the quality of student bibliographies. Glasgow Caledonian University shared the same project objectives around information literacy as Leeds Metropolitan University. Surveys of current students and alumni were carried out. The study clearly demonstrates the difficulties of seeking to measure impact: identifying the library's contribution separately, linkages with other variables, and discipline differences in the student population. The study concluded that there was a more sharply focused appreciation of the importance of the information literacy skills, skills taught at university, amongst alumni than current students (Crawford, 2006). This highlights the importance of seeking to measure longer-term impacts.

In another example, the University College Chester demonstrated that, by having librarians working closely with academic staff in making links to high quality e-resources from virtual learning environment (VLE) modules, there was increased use of them. They were able to use this evidence to argue for additional staffing (Payne *et al.*, 2004).

What did the initiative achieve overall? First, and perhaps most important, it has highlighted the importance of seeking to measure impact. Secondly, the Initiative has explored methods for measuring impact. Thirdly, the initiative has built up a network of experience of measuring impact. It has also produced examples and tools that can be more widely shared. Participants have felt that the materials and experience could be built upon to support the creation of a "community of practice". Finally, the initiative

has also promoted practitioner-based research. A total 22 institutions have been involved in projects that they may not otherwise have undertaken.

The initiative has confirmed that it is not easy to assess impact. This is primarily because we are usually dealing with assessing the impact on people – changes in their behaviour, knowledge, awareness, competencies, and attitudes. Not only are these things difficult to measure, but it is often a challenge to disentangle our contribution to the change from the contribution of others.

The initiative was not without difficulties. Participants at the end of phase 2 noted several problem areas. First, it was often difficult to find the time to undertake the work. Most participants reported that it took more staff time than they had anticipated. Some found that pressing operational priorities interfered with their ability to conduct their impact study. Secondly, participants felt that there was a need to encourage greater sharing of experience between participating institutions. There were a couple of examples of collaborative projects across the two phases of the initiative but this might have been encouraged more. Thirdly, in view of the more extensive of "softer" methods, participants felt that they needed more assistance in the use of qualitative research methods and in the analysis of qualitative research data. Fourthly, participants felt that there was a need to take greater account of prior experience of conducting research and adjust the support accordingly. Fifthly, participants would have liked more ongoing support between events. This included a more pro-active approach to nudging and encouraging the project teams. Finally, participants felt that there was a need for greater clarity at the outset of what was expected in terms of participation.

However, participating institutions did find that attempting to assess their impact is worthwhile. Managers discovered that undertaking this work has helped managers and practitioners by providing evidence of the effects of new services or innovations. This supports the management of change. But, it goes further. Seeking to assess impact moves us from traditional views of service quality, based on such things as use statistics and satisfaction surveys, to looking at the deeper issues associated with our contribution to learning, teaching, and research.

References

Aabø, S. (2005), "The value of public libraries in Norway", *Bibliotekforum*, Vol. 6, pp. 38-40.

Ahtola, A.A. (2002), "How to evaluate and measure the impact of the library's collection on the learning outcome?", paper presented at the 68th IFLA, Glasgow, available at: www.ifla. org/VII/s2/conf/ahtola.pdf (accessed 11 February 2006).

Association of College and Research Libraries (ACRL) (1998), "Task force on academic library outcomes assessment report", available at: www.ala.org/ala/acrl/acrlpubs/whitepapers/ taskforceacademic.htm (accessed 11 February 2006).

Beile, P.M., Boote, D.N. and Killingsworth, E.K. (2004), "A microscope or a mirror? A question of study validity regarding the use of dissertation citation analysis for evaluating research collections", *The Journal of Academic Librarianship*, Vol. 30 No. 5, pp. 347-53.

Blagden, P. (2005), "The LIRG/SCONUL measuring impact initiative overview of phase 1 impact projects", *Library and Information Research*, Vol. 29 No. 91, pp. 14-15, available at: www. cilip.org.uk/specialinterestgroups/bysubject/research/publications/journal/archive/lir91/ 91_blagden.htm (accessed 27 February 2006).

Bohme, S. and Spiller, D. (Eds.) (1999), *Perspectives of Public Library Use 2: A Compendium of Survey Information*, Library and Information Statistics Unit (LISU), Loughborough.

Brown, C. and Krumholz, L.R. (2002), "Integrating information literacy into the science curriculum", *College and Research Libraries*, Vol. 63, pp. 111-23.

Cram, J. (2000), "Six impossible things before breakfast: a multidimensional approach to measuring the value of libraries", in *Proceedings of the 3rd Northumbria International Conference on Performance Measurement in Libraries and Information Services, University of Northumbria, Newcastle upon Tyne*, pp. 19-29, available at: www.alia.org. au/~jcram/six_things.html (accessed 11 February 2006).

Crawford, J. (2006), "The use of electronic services and information literacy: a Glasgow Caledonian University study", *Journal of Librarianship and Information Science*, Vol. 38 No. 1, pp. 33-44.

De Jager, K. (1997), "Library use and academic achievement", *South African Journal of Library and Information Science*, Vol. 65, pp. 26-30.

De Jager, K. (2001), "Impacts and outcomes: searching for the most elusive indicators of academic library performance", in *Meaningful Measures for Emerging Realities, Proceedings of the 4th Northumbria International Conference on Performance Measurement in Libraries and Information Services*, Association of Research Libraries, Washington DC, pp. 291-7, available at: www.arl.org/stats/north/powerpoints/dejager-pt4nor2.ppt (accessed 11 February 2006).

De Jager, K. (2002), "Successful students: does the library make a difference?", *Performance Measurement and Metrics*, Vol. 3, pp. 140-4, available at: http://zerlina.emeraldinsight. com/pdfs/awards2003/pmm2.pdf (accessed 11 February 2006).

Debono, B. (2002), "Assessing the social impact of public libraries: what the literature is saying", *Australasian Public Libraries and Information Services*, Vol. 15 No. 2, pp. 80-95.

Dunn, K. (2002), "Assessing information literacy skills in the California State University: a progress report", *The Journal of Academic Librarianship*, Vol. 28, pp. 26-35.

Everest, K. and Payne, P. (2001), "The impact of libraries on learning, teaching, and research: report of the LIRG seminar held in Leeds 30th, October 2001", *Library and Information Research*, Vol. 25 No. 81, , pp. 18-22, available at: http://eprints.rclis.org/archive/00003337/ 01/article81c.pdf (accessed 20 April 2001).

Everest, K., Scopes, M. and Taylor, J. (2005), "Leeds Metropolitan University: the impact of the information literacy framework", *Library and Information Research*, Vol. 29 No. 91, pp. 26-7, available at: www.cilip.org.uk/specialinterestgroups/bysubject/research/ publications/journal/archive/lir91/91_everest_scopes_taylor.htm (accessed 27 February 2006).

Fister, B. (2003), "What do they know? Assessing the library's contribution to student learning", *Library Issues*, Vol. 23 No. 3.

Hess, F.M. and Klekotka, A.L. (2005), "Impact/outcome evaluation", *Encyclopedia of Social Measurement*, Vol. 2, pp. 271-6.

Holt, G.E., Elliott, D. and Moore, A. (1999), "Placing a value on public library services", *Public Libraries*, Vol. 38, pp. 89-108.

International Federation of Library Associations and Institutions (IFLA) (2005), "Alexandria manifesto on libraries: the information society in action", available at: www.ifla.org/III/ wsis/AlexandriaManifesto.html (accessed 11 February 2006).

Kayß, M. and Poll, R. (2006), "Unterstützen Bibliotheksbestände die Forschung? Zitatanalyse in Dissertationen", *B.I.T. Online*, No. 2.

Linley, R. and Usherwood, B. (1998), *New Measures for the New Library: A Social Audit of Public Libraries*, British Library Board, London, available at: www.ifla.org/IV/ifla64/054-94e.htm (accessed 11 February 2006).

Mark, A.E. (2004), "Assessing student learning: available resources", *College and Research Libraries News*, Vol. 65, pp. 254-6, available at: www.ala.org/ala/acrl/acrlpubs/crlnews/backissues2004/may04/studentlearning.htm (accessed 11 February 2006).

Markless, S. and Streatfield, D. (2005), "Facilitating the impact implementation programme", *Library and Information Research*, Vol. 29 No. 91, pp. 8-13, available at: www.cilip.org.uk/specialinterestgroups/bysubject/research/publications/journal/archive/lir91/91_markless_streatfield.htm (accessed 27 February 2006).

Markless, S. and Streatfield, D. (2006), "Gathering and applying evidence of the impact of UK university libraries on student learning and research: a facilitated action research approach", *International Journal of Information Management*, Vol. 26, pp. 3-15.

Middleton, A. (2005), "An attempt to quantify the quality of student bibliographies", *Performance Measurement and Metrics*, Vol. 6 No. 1, pp. 7-18.

Morris, A., Hawkins, M. and Sumsion, J. (2001), *The Economic Value of Public Libraries, Resource: The Council for Museums*, Archives and Libraries, London.

Payne, P. and Conyers, A. (2005), "Measuring the impact of higher education libraries", *Library and Information Research*, Vol. 29 No. 91, pp. 3-7, available at: www.cilip.org.uk/specialinterestgroups/bysubject/research/publications/journal/archive/lir91/91_payne_conyers.htm (accessed 27 February 2006).

Payne, P., Crawford, C. and Fiander, W. (2004), "Counting on making a difference: assessing our impact", *VINE: The Journal of Information and Knowledge Management Systems*, Vol. 34 No. 4, pp. 176-83.

Pung, C., Clarke, A. and Patton, L. (2004), "Measuring the economic impact of the British Library", *New Review of Academic Librarianship*, Vol. 10 No. 1, pp. 79-102.

Revill, D. (1990), "Performance measures for academic libraries", in Kent, E. (Ed.), *Encyclopedia of Library and Information Science*, Dekker, New York, NY, pp. 294-333.

Smith, E.T. (2003), "Assessing collection usefulness: an investigation of library ownership of the resources graduate students use", *College and Research Libraries*, Vol. 64 No. 5, pp. 344-55.

Tuñon, J. and Brydges, B. (2005), "Improving the quality of university libraries through citation mining and analysis using two new dissertation bibliometric assessment tools", paper presented at the 71st IFLA, Oslo, available at: www.ifla.org/IV/ifla71/papers/078e-Tunon_Brydges.pdf (accessed 11 February 2006).

Corresponding author
Roswitha Poll can be contacted at: pollr@uni-muenster.de

Publishing solutions for contemporary scholars: the library as innovator and partner

Sarah E. Thomas

Cornell University Library, Ithaca, New York, USA

Abstract

Purpose – The purpose of this article is to review the trend in academic libraries toward including scholarly communication, and by extension, electronic publishing, as part of their core mission, using the Cornell University Library as an example.

Design/methodology/approach – The paper describes several manifestations of publishing activity organized under the Library's Center for Innovative Publishing, including the arXiv (http://arxiv.org/), Project Euclid (http://projecteuclid.org), and DPubS (http://DPubS.org).

Findings – Libraries bring many competencies to the scholarly communications process, including expertise in digital initiatives, close connections with authors and readers, and a commitment to preservation. To add publishing to their responsibilities, they need to develop expertise in content acquisition, editorial management, contract negotiation, marketing, and subscription management.

Originality/value – Academic libraries are making formal and informal publishing a part of their core activity. A variety of models exist. The Cornell University Library has created a framework for supporting publishing called the Center for Innovative Publishing, and through it supports a successful open access repository (arXiv), a sustainable webhosting service for journals in math and statistics (Project Euclid) and a content management tool (DPubS) to enable other institutions (libraries, scholarly societies, presses) to engage in similar ventures to increase the dissemination of scholarship and to lower the barriers to its access.

Keywords Academic libraries, Electronic publishing

Paper type Case study

Introduction

What can an academic library contribute to scholarly publishing? The Cornell University Library has engaged in a number of activities in the publishing realm that aim at increasing affordable, effective, widespread, and durable access to research. The Library's Center for Innovative Publishing (CIP) operates the arXiv[1], an e-print service for physicists, computer scientists, mathematicians, and others; Project Euclid[2], a journal hosting service for over 40 titles in math and statistics; and is developing, with Pennsylvania State University, DPubS[3], an open source publications management software. The library also runs an implementation of DSpace. Cornell's DCAPS, or Digital Consulting and Production Service, assists in the transition of print to electronic through its digitization, metadata production, and consulting service. Digital publications are preserved according to a well-developed policy for digital archiving, ensuring ongoing access to information across time.

This article was previously published in *Publishing Research Quarterly*, Vol. 22 No. 2, Summer 2006.

The Cornell University Library's Center for Innovative Publishing is one manifestation of publishing activity undertaken by academic libraries as part of a movement to increase access to scholarship in an affordable manner and to ensure the ongoing availability of scholarly information in a way that is consistent with the traditional library role of preserving the record of our civilization from generation to generation. CIP also seeks to apply innovative techniques in the management and delivery of information to scholars.

Publishing activity undertaken by libraries is today often viewed as innovative. Some individuals have even questioned whether publishing is an appropriate role for the library. Mission statements for libraries in the pre-digital age tended to focus on the librarian's role in working with the end products of scholarship. For example, this 1981 (reaffirmed 1993) mission statement is typical of many academic libraries of the recent past:

> The mission of the University of Delaware Library is to gather, organize, preserve, and provide access to the information resources necessary for the University of Delaware to achieve its educational, research, and service goals (University of Delaware Library, 1993).

The function of publishing was not commonly included in the mission of libraries of the twentieth century. Yet in the nineteenth century, there were examples of close relationships between libraries and publishing. At Cornell, the first university librarian, D. Willard Fiske, appointed in 1868, also served, beginning in 1869, as the first director of the university press.

In the first decade of the twenty-first century libraries are reframing their mission to reflect changes in their environment and the expectations of their users. At the forefront of their strategic priorities it is common to find objectives relating to publishing and scholarly communications. One recent research library mission statement proclaims:

> The mission of the MIT libraries is to create and sustain an intuitive, trusted information environment that enables learning and the advancement of knowledge at MIT. We are committed to developing strategies and systems that promote discovery and facilitate worldwide scholarly communication (MIT Libraries, 2003).

At the Association of Research Libraries, a North American organization with 123 member libraries, strategic planning undertaken in 2004 has resulted in three priorities: scholarly communication; public policies affecting research libraries; and research, teaching, and learning. In the last decade organizations such as SPARC have evolved to promote more cost-effective dissemination of scholarly work. The Scholarly Publishing and Academic Resources Coalition (SPARC) is supported by over 213 members, including a broad cross section of college and research libraries. And in another sign of how publishing is increasingly seen as the purview of librarians, a growing number of library directors oversee the university press at their institution. Recent examples of convergent administration of libraries and university presses include the Massachusetts Institute of Technology, New York University, Northwestern University, Penn State University, and Stanford University. Many more research libraries now are engaging in the dissemination of scholarly information, both formally and informally. University libraries are lending their digital expertise and information technology infrastructure to host online journals and dozens of North American academic libraries have created institutional repositories to collect and disseminate the research and courseware of their faculty. Brigham Young

University's library publishes 12 online journals, for example, and the California Digital Library has made available online over 2000 books through its eScholarship Editions. In February 2005 a survey of the membership of the Coalition for Networked Information found that about 40 percent of respondents had an operational institutional repository (IR), and that 88 percent were planning an IR or participation in a consortial IR (Lynch and Lippincott, 2005).

Early publishing activities and influences

At the Cornell University Library, the immediate antecedents for its contemporary publishing activities were an outgrowth of its exploration of emerging imaging technologies and a response to financial pressures created by the so-called "scholarly communications crisis." In the late 1980s, Cornell, in partnership with Xerox Corporation, established a pioneering digital imaging project, which scanned almost 600 out-of print monographs in mathematics. This republishing enterprise prefigured Google's mass digitization and has remained vital to this day, with continuous use of the collection. The library provides print-on-demand for titles in the Cornell math books[4]. Ongoing reformatting of print collections into digital versions has continued, with an expanding list of titles numbering in the thousands. Among the largest of the Cornell digitization projects are the Core Historical Literature of Agriculture[5], Home Economics Archive: Research, Tradition, History (HEARTH)[6], and the Making of America[7]. Without realizing it, the Cornell University Library and others like it had become informal publishers of retrospective materials.

Simultaneously in the late 1980s and early 1990s librarians became more vocal about the rising prices of journals, and cast about for solutions to the so-called "serials crisis." By the 1990s several partnerships between librarians and publishers had emerged. The goals of these collaborations were to support publication by the academy for the academy and to turn the tide in the pricing situation. Often funded with start-up subsidies from foundations, these online services offered an alternative to profit-driven commercial journals or stated as their objective the intention of moderating price increases. Project Muse, begun at Johns Hopkins in 1993 as a collaborative endeavor of the university libraries and the Johns Hopkins University Press, and funded by the Andrew W. Mellon Foundation and the National Endowment for the Humanities, bundled many titles, primarily in the humanities and social sciences with the goal of bringing the full text of high quality scholarship online in a sustainable manner. Their initiative has helped over 300 journals, products of scholarly societies and other academic organizations, make the transition from print to electronic by sharing an infrastructure and common access. Another library-associated publishing model is offered by HighWire Press, a division of the Stanford University Libraries. Since 1995 HighWire has provided webhosting services for publishers and has sought to improve the environment for users of scholarly information by advocating open access for backfiles:

> HighWire was founded to ensure that its partners – scientific societies and responsible publishers – would remain strong and able to lead the transition toward use of new technologies for scientific communication. Concerned that scientific societies separately would lack the resources and expertise to lead a major technical infrastructure shift in publications, Stanford University, in founding HighWire, accepted the role of partner, agent of change, and advisor. Begun as a close collaboration of scientists, librarians and publishers, it has not strayed from that model in its six years of rapid growth (HighWire Press, 2001).

Project Euclid

At Cornell, influenced by these precursors and encouraged by mathematicians on the faculty with a drive to convert the literature of their discipline to online form, the library received a grant from The Andrew W. Mellon Foundation in 1999 to create a service to deliver journals in math and statistics electronically. The Mellon Foundation provided a second grant in 2003 for a total investment of $1,250,000. The grant funded Project Euclid, an online publishing platform for math and statistical journals. Project Euclid offered a mixed model of titles sold in a bundle, open access journals, and titles sold separately or available only through society membership. Its business plan called for sustainability based on distributing the costs of operating the journal hosting service between publishers and libraries. Fixed costs for managing Project Euclid in 2005 were approximately $300,000. In 2005 Euclid offered 40 journals and had 200 subscribers, including consortial purchases that made its titles available to every academic institution in India. Publishers selecting Project Euclid as their means of disseminating scholarly work include a mathematics department at a university, a small commercial firm, several scholarly societies, and a university press. Cornell has seen its principal market as small, independent journals, of which there are many in the field. Two of the top titles in mathematics appear in the list, which is international in character.

From the outset, Project Euclid sought to capitalize on the synergy created by bringing related titles together. Ability to search across the full text of those files and to follow links from MathReviews to Euclid texts or to trace a reference from an article in Euclid to a related publication is an essential aspect of Euclid's design. Project Euclid has received positive reviews and has won the Charleston Advisor's Reader's Choice award for best pricing for its reasonable and diverse models of pricing. As of the fourth quarter of 2005, Project Euclid is in the black, having achieved sustainability within three years of its public launch. In the process the Cornell University Library has learned a great deal about acting as a digital press. As a start-up, the library needed to build confidence among publishers that it could function as a press. Although the library's brand was strong in the information and digital innovation areas, it had little or no recognition as a press. Publishers and scholarly societies were skittish about trusting their journals with an unproven partner, especially since those journals contributed considerable to their reputations and their bottom line. There were significant aspects of publishing where the Library had scant experience, such as marketing or handling subscription requests. And in defining the scope of Euclid, Cornell introduced a complexity that made for a challenging business model.

In 1999, when Euclid was conceived, its models were Project Muse and HighWire. Open access was not yet deeply rooted. The math environment proved highly diverse. In order to satisfy the varying needs of Project Euclid's first partners, the initiative needed to accommodate publishers who wanted their journal to be open only to members of their society, journals that would be hosted by Euclid but available only through the publisher, journals willing to be part of a bundle of federated titles that would be sold to subsidize the online service, and journals that wanted to be freely available to all. This made for a complicated pricing structure for publishers. Cornell sought to create the best value through an aggregation. The environment that had drawn Cornell to pursue mathematics as a disciplinary focus also increased the cost of the effort. The motivations for Project Euclid are to increase access to scholarship, offer economies of scale, and to

provide an alternative for publishers who might otherwise be enticed to sign on to a commercial contract that would be favorable for their publication, but costly for institutional subscribers. The mathematics journal market was large, with over 500 journals considered core by one of the most prominent indexing services, MathReviews. These 500 journals were published by hundreds of publishers, many offering one, two or three titles only. But the fragmented array of publishers meant the Project Euclid staff had to interact with many different publishers, often a time-consuming process. Furthermore, since some of these publishers operated on very tight margins and had little experience with online access, agreeing to participate in Euclid required a leap of faith along with a commitment to "do the right thing" – make their publications available in an affordable manner.

Project Euclid operated under a subsidy from the Andrew W. Mellon Foundation since 1999, when it received development funds, through its launch in May 2004 until November 2005, when revenues began to exceed operating expenses. Its current business model balances fees paid by publishers with revenue from subscriptions. One of the startling insights of running a digital press service on a cost-recovery basis was that the library had a much deeper sympathy with publisher pricing models than it had had when it was only a consumer of publishers' products. It had to walk the line between pricing that the publishers could afford and licenses that would be library-friendly. The need to develop a successful business model underscored an area where the library lacked substantial expertise, that of cost assessment and analysis. Since developing Euclid as an enterprise, the library has added another master of business administration (MBA) to its staff, and several other staff members are enrolled in an executive MBA program. The library also was unfamiliar with the process of acquiring new content, with producing print publications beyond a boutique scale, with marketing, and with managing subscription access or fulfillment. It has outsourced some of these functions, but as Euclid matures, it is increasingly bringing them in house, but hiring experienced staff to manage them.

Other publishing competencies the library has had to develop and is still assimilating are the knowledge of editorial management procedures and the ability to negotiate contracts with journal owners. Libraries also have limited background in working with printers. However, the library is entering publishing at a time when many of its processes are undergoing significant change. Publications are transitioning to e-only; print on demand and short run printing, made possible through evolving technology, are altering traditional print practice. Amazon and Google, with services such as Book Surge, GooglePrint, and GoogleLibrary, are disruptive forces in the publishing environment. Consequently, both publishers and librarians need to develop new skills and apply them to the rapidly changing arena of scholarly communications.

Now three years since its public launch, Project Euclid is healthy and growing, with four journals poised to come online in spring 2006 and with backfiles being converted to expand its depth. It continues to build on its original strengths, including the library's ability to execute complex digitization projects, create metadata, and serve users 24/7/365. The library is also contributing its expertise in digital preservation. We expect Project Euclid to expand both in numbers of titles offered and in the number of users. Still, scholarly journal publication is fluid, and we can expect changes in the coming decade. Two-thirds of Project Euclid's 36,000 articles are open access. Will small publishers be able to continue amidst the financial pressures that beset them?

Will the American Mathematical Society, a prominent and respected mathematical society and publisher, offer a service that will unite hundreds of math journals under one umbrella without overshadowing the smaller, independent societies and departments making their titles available through Euclid? Sustainability is a moving target in publishing, and as the Cornell University Library becomes more heavily invested in this function, it is also exploring other, less traditional and less formal, approaches to publishing.

arXiv

Another publishing activity now housed in the Cornell University Library is the arXiv, the e-print archive originally established to support the online exchange of preprints in high energy physics. An example of an alternative mode of publishing, the arXiv was begun on a shoestring budget in 1991 by Paul Ginsparg, then a scientist at the Los Alamos National Laboratory. When Ginsparg joined Cornell as a member of its faculty in 2001, the library made a strong pitch that this dynamic information resource should become part of its ongoing operations. The arXiv remains today the most successful open access repository in the world. Submissions have increased over 60 percent since 2001, averaged 4,000 per month in 2005, and continue to rise. Use is intense. At Cornell we track about 300 million hits per year, and with several mirror sites around the world, this use is but a fraction of the total. The arXiv has introduced new functionality into the world of physics publishing, providing a low-cost, rapid means of disseminating scholarship that has transformed communication among the physics community. Expanding from high-energy physics, the arXiv now takes in submissions in almost equal amounts in high-energy physics, condensed matter physics, astrophysics, and mathematics. The e-prints coexist with formal publications. Although the working scientist relies on the arXiv for up-to-date, almost comprehensive access to research in physics and related disciplines, she continues to submit her work for review and publication in formal journals. Libraries continue to subscribe to these journals. Formal publication remains vital for reputation, promotion and tenure, and preservation purposes. The cost of the added functionality provided through the arXiv is slight, currently averaging about $4 per article. In contrast, a peer-reviewed, formally published article might range from $1,500 to $2,500. The library received an initial three-year subvention from the university provost to fund the transition of the arXiv from federal support provided by the National Science Foundation and the US Department of Energy to operations at Cornell, and since FY 2004/2005 the Cornell Library has reallocated approximately $200,000 annually from its budget to support the day to day operations of the arXiv. The Library provides the lion's share of support for daily operations, while development of new directions and research continues to come from grants and external subsidies. One of the key ways in which the library expects to contribute its unique expertise to the established success of the arXiv is by ensuring its preservation. The library is developing an open archival information system (OAIS) based on international standards and capable of ensuring long-term preservation of digital content.

The arXiv, like other publishing activities, is evolving. It is maturing into a service with a more polished interface and now has an advisory board and named moderators. Efforts are underway to facilitate the use of arXiv submissions by publishers, enabling them to have more efficient workflows. There are subdisciplines that are good

candidates for coverage in the arXiv, so the scope of the database may expand. Within the library, we hope to increase the synergy among the platforms and software used to support the various publishing initiatives overseen by the Center for Innovative Publishing. The relationship between the contributors to the arXiv and their publishers is quite strong, and it appears that for now, the value added by the traditional publisher, often their professional society, is strong enough that forgoing the final peer-reviewed version would be unthinkable for most authors. In 2005, however, ISI began indexing deposits in the arXiv and other open access resources in its web citation index, with the possible consequence that the impact of informal publications can be measured alongside traditional formal peer-reviewed publications. One of the outcomes to watch for would be a willingness to rely on the informal peer review that characterizes the arXiv and to endorse a new means of assessment such as citation to works in the arXiv that might diminish the need to subscribe to journals which have major overlap with arXiv submissions.

DPubS

The third significant publishing initiative of the Cornell University Library is the creation of DPubS, an open source content management program. In 2004, as institutional repositories began to ascend in prominence in the US, there were observations that they would benefit from additional functionality. They served effectively as a means for scholars at an institution to deposit their intellectual efforts, and thus as a foundation for the preservation of those contributions by the library. However, institutional repositories have not substituted for traditional publications, and thus have not had a substantial impact on the journals pricing situation.

At Cornell, Robert Cooke, a faculty member, received a grant from the Atlantic Philanthropies that supported the Internet-First University Press. Through this initiative, the library implemented DSpace. DSpace, a joint effort of the Massachusetts Institute of Technology and Hewlett Packard, is open source software to support digital repositories. The DSpace site lists almost 150 installations worldwide, and many more libraries and other organizations have experimented with the easily downloaded software. To increase the access capability of DSpace, Cornell plans to combine it with an enhanced version of the content management software it is using to support Project Euclid. It was a logical step to add functionality to DSpace and to join the burgeoning open source movement. Cornell named the generalized version of the software DPubS to distinguish the tool from its application in Project Euclid. In addition, the library also decided to program interoperability with Fedora, an open source software that provides an architecture for managing and delivering digital content. Developed in collaboration by the University of Virginia Library and Cornell University's Digital Library research Group in Information Science, Fedora is being used by a growing community of institutions supporting scholarly research. The library anticipates that the ready availability of the DPubS content management tool will enable more institutions to undertake both informal publication through IR's and more formal publication, such as the type supported in Project Euclid.

At the same time the Cornell University Library was expanding its use of DPubS locally, the Pennsylvania State Libraries and the Penn State Press were beginning a collaboration in the Office of Scholarly Publishing. They intended to acquire content management software to support journals, conference proceedings, and other

documents and to begin their concentration with support for history, particularly the history of Pennsylvania and the Northeast, romance studies, and nutrition. Their search for software had not turned up a product that met all their requirements. One important criterion was the ability to be able to access the code directly for local modification, and many of the options available to them were proprietary. Penn State's dean of libraries offered to work with Cornell to develop a version of Euclid that could be imported by Penn State. Out of this emerged a joint proposal to the Andrew W. Mellon Foundation to generalize and extend the journal management software underlying Project Euclid. The partners proposed to extend the formats supported, add editorial management capability, and to make the product open source. One of their goals was to create a flexible tool that would increase the sustainability of the Penn State Press by enabling it to support a variety of economical but revenue-generating online publications. Another goal was to increase the dissemination and access to scholarship that had limited commercial value but that was important for the advancement of knowledge. Conference proceedings often fell into this category. Thirdly, both institutions saw the DPubS software as the means of offering an alternative venue for the publication of academic works that were being submitted for publication in journals managed by for-profit entities. As the literature has documented, some commercial publishers and even some societies promote journals and other works that cost significantly more per page than some other well-respected titles (Bergstrom and Bergstrom, 2005). Although these studies have been controversial, many librarians have concluded the alternatives to the present model would be beneficial to the budgets of libraries and universities. They reasoned that open access publications or low cost titles would be feasible if the cycle of knowledge creation and dissemination were contained within the academy and its close collaborators, such as scholarly presses and societies (Suber, 2003).

Conclusion

The first decade of the twenty-first century is rife with debate about the merits of open access and the prognosis for the continuation of traditional journals. Within Cornell University itself there is no single position with regard to the future. A library task force on open access produced a report in 2005 that calculated that a complete transition to an author-pays model would result in a research-intensive institution such as Cornell paying more than its current allocation for subscriptions (Davis *et al.*, 2004). Critics of the report have noted that the amount used as the basis for its estimate, that of a cost of $2,500 per article, lies on the high end of article costs, and that a figure of $1,500 is more appropriate (Suber, 2005). Further, they have suggested that the task force did not include savings from costs currently borne outside the library, such as page charges, that would mitigate the total cost to the university. Finally, they raise the issue of cost-benefit. The advantage to the scientists and scholars of having their work freely accessible throughout the world, with the consequence of accelerated transmission of ideas, creation of new knowledge, and enhanced influence and reputation for authors is worth the investment, even if an open access model should prove more expensive for a large research institution such as Cornell.

Bound up in this debate are a host of other concerns that color the discussion around open access. Many faculty confuse open access with unmoderated lists, and think that open access journals are not peer-reviewed or are lesser quality. Recent studies of the

impact factor of open access journals and increased publicity about the value of open access are beginning to penetrate faculty circles, but in general there is not a solid understanding of the issues. In addition, the role of journals in the financial sustainability of scholarly societies creates a conflict in members who, while sympathetic to the idea of broader dissemination of scholarship and open access, see other programs of their society, such as training of entering scholars, conferences, and even the existence of their organization itself, threatened by the loss of the cross subsidy that their publications operations provide. The disconnection between the producers of scholarly literature and the intermediaries who purchase it for consumption by others has generated a dysfunctional economic relationship.

At the same time, emerging technologies have created another destabilizing or challenging environment for publishers. To meet the demand for online access and to take advantage of new information technologies that provide laborsaving, time saving, and intellectual enhancements, publishers have invested heavily in building repositories and interfaces and in converting backfiles. They have grappled with new business models that take into account the loss of print subscriptions to e-only access. They have replaced subscriptions with database contracts. Publishers who were accustomed to interacting with intermediaries, vendors who had relationships with another set of intermediaries, librarians, are now finding themselves marketing directly to libraries or even to consumers, resulting in a considerably different service environment.

In recognition of the reality of this situation and the time that will be needed to deconstruct it and to create a new, healthier system of academic exchange, DPubS is a tool that supports a variety of business models of scholarly communication. One of its key assets is its access control mechanism. A scholarly society can employ DPubS to offer a complimentary subscription to its journal for members and fee-based access for non-members. An organization can subsidize the dissemination of research in a discipline, resulting in an open access publication. The Cornell University Library and the Institute for Advanced Study at Princeton University are collaborating in making the *Annals of Mathematics*, the world's top impact journal in mathematics, freely available through Project Euclid. University presses can shore up their financial foundations through the successfully proven method of selling journal subscriptions, but still price their products at a lower cost than large for-profit entities. The hope is, with the technical infrastructure provided by library collaborators and the flexible tool available in DPubS, that university presses will be a positive alternative publisher of scholarly journals. Institutional repositories, a growing trend in higher education, both to collect and preserve the intellectual output and diverse learning objects of universities and to increase access to and flexible use of scholarly materials, can also use DPubS to advantage. Federated institutional repositories can create overlay journals or disciplinary groupings using the DPubS software. One can imagine distributed but interconnected centers of excellence that link scholarship in various subject domains: labor history, nanofabrication, Islamic studies, philosophy, and others. Depending on the financial model chosen to support the dissemination and exchange of scholarly information, organizations can control access using DPubS.

The Cornell University Library has designated as two of its top priorities increasing the understanding of the Cornell community of issues in scholarly communications and promoting e-publishing solutions that benefit higher education. In practical terms, this

means reaching out to faculty and graduate students through an educational campaign that urges authors to "know your publisher," "manage your copyright," and "store your stuff." Behind these exhortations are details of the relative costs of journals and their impact factors, information about copyright and fair use, and background about preservation in an increasingly digital world. The library offers, through its Digital Consulting and Production Service[8], a one-stop shopping service with connections to digital imaging experts, metadata services, copyright clearance and advice, archiving, and publishing. Informed faculty partners and a well-designed enabling infrastructure are prerequisites to successful library publishing endeavors.

As the scholarly enterprise becomes more complex, librarians are finding their expertise is frequently called upon earlier in the process of research and teaching. In the past, some academics viewed libraries as fairly passive organizations entrusted with the stewardship of books and journals consulted by faculty in their research and students in their learning. An evolving model is more collaborative and interactive. Researchers work in cross-disciplinary teams, and students engage in collaborative learning experiences. Librarians are contributing their knowledge of information management, organization, and sources more actively to the educational endeavor, and they are more likely to become a member of a faculty production team that is creating a dynamic learning object. With their experience in the digital domain and their familiarity with a broad spectrum of the end products of research, scholarly publications, they are well placed to facilitate innovative models of scholarly communication. By engaging in publishing and collaborating with authors, scholarly societies, computer scientists, and other stakeholders to create a twenty-first century process for communicating and using scholarly research, librarians are pursuing important goals: the reduction of the costs to the academy and society of sharing scholarly information and the lowering of the barriers to access to the knowledge created within the university.

The Cornell University Library's exploration of a variety of publishing ventures has matured into a service organization known as the Center for Innovative Publishing. This enterprise embraces a self-supporting and growing online service for the federation of journals in mathematics and statistics known as Project Euclid; a subsidized open access disciplinary repository of world-wide renown in the physics, math, and computer science communities, the arXiv; an institutional repository, DSpace; flexible open source content management software, DPubS, that will launch publicly in summer 2006; and a service bureau for publishing content in all disciplines and in a variety of formats. With the coalescence of the Center for Innovative Publishing into a single unit, the library has moved from a collection of grant-funded or cost-recovery projects into a solid program that will be integrated into its financial base and that will support the core mission of libraries, creating and sustaining a trusted information environment and developing strategies and systems that promote discovery and facilitate worldwide scholarly communication.

Notes
1. http://arXiv.org
2. http://projecteuclid.org
3. http://DPubS.org

4. http://historical.library.cornell.edu/math/
5. http://chla.library.cornell.edu/
6. http://hearth.library.cornell.edu/
7. http://cdl.library.cornell.edu/moa/
8. http://dcaps.library.cornell.edu/

References

Bergstrom, C.T. and Bergstrom, T.C. (2005), "The economics of scholarly journal publishing", available at: http://octavia.zoology.washington.edu/publishing/ (accessed February 26, 2006).

Davis, P.M., Ehling, T., Habicht, O., How, S., Saylor, J. and Walker, K. (2004), "Report of the CUL task force on open access publishing presented to the Cornell University Library management team, August 9, 2004", available at: http://dspace.library.cornell.edu/handle/1813/193 (accessed February 26, 2006).

HighWire Press (2001), available at: http://web.archive.org/web/20010616124819/www.highwire.org/intro.dtl (accessed February 15, 2006).

Lynch, C. and Lippincott, J.K. (2005), "Institutional repository deployment in the United States as of early 2005", *D-Lib Magazine*, Vol. 11 No. 9, available at: www.dlib.org/dlib/september05/lynch/09lynch.html (accessed February 24, 2006).

MIT Libraries (2003), "Mission", available at: http://libraries.mit.edu/about/mission.html (accessed March 29, 2006).

Suber, P. (2003), "Removing the barriers to research: an introduction to open access for librarians", available at: www.earlham.edu/~peters/writing/acrl.htm (accessed February 26, 2006).

Suber, P. (2005), SOAN April 2, 2005, available at: www.earlham.edu/~peters/fos/newsletter/04-02-05.htm (accessed February 26, 2006).

University of Delaware Library (1993), "Mission statement", available at: http://www2.lib.udel.edu/policies/mission.htm (accessed March 29, 2006).

Corresponding author
Sarah E. Thomas can be contacted at: set9@cornell.edu

Data for the future

The German project "Co-operative development of a long-term digital information archive" (kopal)

Reinhard Altenhöner

Die Deutsche Bibliothek (DDB), Frankfurt am Main, Germany

Abstract

Purpose – One of the unresolved problems of the global information society is ensuring the long-term accessibility of digital documents. The project kopal tackles this problem head-on: in a three-year project kopal's objective is the practical testing and implementation of a cooperatively created and operated long-term archival system for digital resources.

Design/methodology/approach – The system will be implemented in accordance with international standards for long-term archiving and metadata within the Open Archival Information System (OAIS) framework. The project partners, Die Deutsche Bibliothek (DDB), Göttingen State and University Library (SUB Göttingen), IBM Deutschland GmbH and the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), will establish a cooperatively transferable solution for cultural heritage institutions, as well as for business and industry.

Findings – Within the project, the project partners DDB and SUB Göttingen are developing software for the input and access of data, which will be released under an open-source license.

Research limitations/implications – Long-term preservation methods and strategies will be discussed in general in the paper.

Practical implications – The project will present a stable and reusable platform for additional partners and users, especially for cultural heritage organisations.

Originality/value – The solution is based on Digital Information and Archiving System (DIAS), jointly devised by IBM and the National Library of The Netherlands in The Hague, and it will be adapted to the needs of the project with several extensions. Establishing a collaborative solution for long-term preservation is a milestone in the development of systems for the long-term availability of digital objects.

Keywords Digital libraries, Archives management, Digital storage, Germany

Paper type Research paper

Publication process

Today, the production phase of the publication process is completely electronic-based, but increasingly the same is true for the delivery of publications to the end-users, the readers. But we still do not know how digital publications, works of art, image and sound documents, files, primary data, and music can be archived so that they will remain permanently readable and thus accessible to all in the future. What we can say at the present time is: whatever strategy is followed in the future to provide access to digital content, it will depend on the existence of a bit stream, the integrity and authenticity of which has been kept in order over the years and decades. So bit stream-preservation has the potential facility to make digitally stored objects available for a long time. Based on this, document rendering will have to be enabled for access to

digital objects in the future. Several strategies are currently under discussion. The most important action points can be summarised in these two bullets:

From today's perspective, migrating electronic objects in a controlled environment from one format to another, will be more usual and better for long-term access.

The alternative is to emulate the historic system environment from the production time of the object – including the emulation of hardware and basic system software.

In addition to the availability of the bit stream, we need metadata information, in particular concerning technical information about the stored object, its original technical environment and its storage and migration history. The task of establishing a system and an infrastructure for long-term preservation is complex and demands a large amount of money and personal resources, and we need an academic network and knowledge-sharing – nationally and internationally.

Die Deutsche Bibliothek

Die Deutsche Bibliothek (DDB)[1] is the national library and national bibliographic information centre for the Federal Republic of Germany. The library is responsible for the collection, processing and bibliographic indexing of all German and German-language publications issued since 1913. This task is based on a statutory mandate for the collection, bibliographic processing and long-term preservation of all publications released in Germany or published in the German language abroad. The law also covers digital publications distributed on physical carriers but makes no provision for online publications until now. Only in the next few weeks (the expectation is that the new law will come into effect in the first half of 2006) will we get new legislation, which will enlarge our area of responsibility to include all types of net or electronic publications. And it is to be expected that in consequence of this new law, a number of regional libraries with legal deposit responsibility for specific regions will get a completely new legal foundation for their collections of digital publications.

Libraries will have to be prepared for this new situation. This is why in the last few years Die Deutsche Bibliothek has started a lot of initiatives to promote the long-term preservation of digital publications in Germany. A number of basic principles applicable to the collection of online publications were defined in preliminary hearings with publishers, library experts, information specialists and government officials and formulated in a policy document passed by the Publishers' Committee of the Börsenverein des Deutschen Buchhandels in June 1997:

- all online publications are to be submitted via data networks or on physical data media upon request;
- online publications available in different forms are to be submitted in the format requested by the library;
- publications with identical contents distributed both on physical media and as online publications are to be submitted in both forms; and
- online publications with identical contents distributed simultaneously by multiple providers need only be submitted once.

Based on these policy principles, Die Deutsche Bibliothek has tested procedures for the submission, collection and long-term preservation of online publications in co-operation with publishers and producers in a test phase lasting several years. In

the process, the "Electronic Deposit Library" taskforce explored and established the conditions necessary for Die Deutsche Bibliothek to become a deposit library for online publications as well.

Since 1998 online dissertations and theses (45,000 so far) have been collected, archived, and made available on a document server. Electronic periodicals have been collected since 2000, and since the year 2001 Die Deutsche Bibliothek has been operating a submission interface for online publications. During the submission procedure, DDB also asks for technical metadata relevant for preservation purposes. This has to be a compromise between the workload acceptable to publishers participating in voluntary submission, and the extensive requirements of the processes in the deposit system for future preservation. Other steps in the coming years are special developments for example for newsletters and for the retrospective ingest of different materials and collections. Furthermore, Die Deutsche Bibliothek has participated in the European Nedlib-project (Networked European Deposit Library) to adopt the open archival information system-model (OAIS) and to develop workflow-suggestions for the integration of library procedures.

Other relevant experiences:

- DDB has built up additional experiences through their system for multimedia access (Multimedia-Bereitstellungssystem/MMB). MMB enables storage and access for digital objects on physical carriers. Different object types (workstation image, application installation kit, file collection, presentation object) have been implemented to provide for the rendering of complex digital objects (applications).

- Another activity covers the development of a persistent identifier infrastructure for Germany. The use of persistent identifiers is the only possibility to guarantee that a digital object can be addressed permanently in the internet. Embedded in the project "EPICUR – Enhancement of Persistent Identifier Services – Comprehensive Method for unequivocal Resource Identification" persistent identifiers become a part of a metadata framework for electronic publications[2]. DDB has chosen the uniform resource name (URN), which is a uniform resource identifier with the term "urn:" preceding the rest of the name, and which serves as a permanent designator for a resource independent of location. URNs are persistent (i.e. they never change) regardless of whether or not the resource's physical location changes. The purpose of URNs is to identify a single resource, and it alone, for the duration of its existence. But it should be noted that a resource can have a number of URNs allocated to it.

National initiatives

Germany has a federal structure with important elements of self-government in the states, especially concerning the education system and the science and research sector. The existence of a lot of regional libraries with legal deposit responsibility for their regions is another part of the federal structure. And considering the importance of the task of long-term preservation within the federal structure of Germany it is obvious that the approach to a successful solution to these issues in Germany must be cooperative. This primarily concerns the organisational aspect, but there are also a lot of practical and technical reasons why we are trying to distribute the responsibility for collecting the electronic objects in the Internet. With this background it is clear that we

need partners in order to implement a long-term preservation strategy and infrastructure in Germany in two directions: organisational and operational.

There are two initiatives – embedded in a lot of smaller projects and initiatives not all of which are mentioned here – through which Germany is trying to approach the problem of long-term preservation:

- From a more general and organisational perspective, network of expertise in long-term storage of online resources (nestor) was established with the goal of building up a platform of competence for sharing knowledge and experience in the field of long-term preservation and to exchange experts and expertise between different types of cultural heritage institutions.

- Die Deutsche Bibliothek, with partners, was given the task of building up a long-term archival system based on OAIS as a practical aspect of the general scope of kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen/Co-operative development of a long-term digital information archive).

nestor

Financed by the German Federal Ministry of Education and Research nestor[3], the alliance for Germany's digital memory, started in 2003 and will be completed in the middle of 2006. Under the leadership of Die Deutsche Bibliothek, there are several partners from the library area (Bavarian State Library, Göttingen State and University Library), media centers (Computer and Media Service of Humboldt University, Berlin), archives (Bavarian State Archives – Head Office) and museums (Institute for Museum Research, Berlin) on board. Additionally, on the advisory board there are publishers, representatives of science and technology, museums, archives, libraries and universities and also members of culture and politics and research institutions / computing centers.

The central aim of the nestor project is to bring together available knowledge, people and expertise on long-term storage of digital resources as a starting point for a future alliance for Germany's digital memory.

Therefore the project:

- creates a network for information and communication about present and future long-term preservation (LTP) activities in Germany;

- establishes a cross-sectoral community to promote and support LTP activities and to raise awareness in society;

- triggers synergies between on-going activities in Germany and cooperates with international partners and projects;

- developes strategies for coordination of LTP activities in Germany; and

- proposes a long-lasting organisational model to continue the service as a network of excellence after the end of project nestor in 2006.

In detail, the tasks and measures are:

- collecting and presenting information;

- consolidating areas of expertise and making them visible and available;

- promoting cooperation and supporting a common solution;

- preparing expertises on technical, organisational and legal issues;
- presenting models, putting them forward for discussion and encouraging widespread best practice;
- promoting standards and representing Germany on international standardisation committees;
- developing collection guidelines and selection procedures for the storage of digital sources;
- coordinating responsibilities for long-term tasks;
- raising awareness of the problem in specialist documentation circles and among the public; and
- preparing a permanent organisation which coordinates and represents the concerns of long-term storage.

An important aspect – mentioned here as an example of the activities – are the working groups and the expert reports. At the moment the following groups are active:

- nestor Working Group on Trusted Repository Certification;
- nestor Working Group on Multimedia Archiving; and
- nestor Working Group on Preservation Policies and Selection Criteria.

The expertises address the following topics:

- electronic Journals;
- perspectives of long-term preservation of multimedia objects;
- development of a descriptive profile for a national long-term preservation strategy (preservation policy);
- digital long-term preservation and the law;
- study of the state of existing research data and raw data from scientific activities;
- a comparison of existing archiving systems; and
- digitisation and preservation of digitised material in German museums.

In the meantime most of them have been published and are available in the internet[4].

kopal

kopal[5], which started in July 2004, faces the problem of long-term preservation from a practical perspective. Technically and organisationally a collaborative approach has been chosen. Financially supported by the German Federal Ministry of Education and Research, kopal is developing an innovative technical solution in the form of a reusable long-term archive for digital data. The solution is – after a market survey done in 2003/2004 – based on Digital Information and Archiving System (DIAS), jointly devised by IBM and the National Library of The Netherlands in The Hague. However, for the purposes of the consortium, there was a need to define some special requirements, especially for the cooperatively and independently usable design of the software solution.

Important points in the requirements and the enhancement of DIAS are:

- remote access capability for partners at different locations;
- flexibility within the system to handle a wide range of formats and different metadata schemes of different (and in comparison to the Netherlands often smaller) partners;
- multi-client capability, handling of personal filestores ("lockers"); and
- separation of DIAS and additional tools for handling ingest and dissemination in order to get a flexible and easily adoptable solution.

The system – like DIAS – is implemented in accordance with international standards for long-term archiving and metadata within the OAIS framework. The possibility of integrating the solution into existing library and information systems is a fundamental objective of the project, and is only possible through transparency by using open and dedicated interfaces.

Some other important complementary components in addition to the existing DIAS-system are:

- realisation of monitoring and steering functions to prepare for the long-term preservation of digital documents (as a starting point to incorporating preservation planning facilities); and
- flexible data import and export functions based on the object description scheme Metadata Encoding and Transmission Standard (METS) of the Library of Congress, USA, in an enhanced and specifically adopted form as universal object format (UOF) (see Figure 1).

The requirements for a universal object format are described as a processing scheme in DIAS Core. In this arrangement, the object is processed as a submission information package (SIP) and delivered as a dissemination information package (DIP), using the OAIS model terminology. Within the DIAS Core, a part of the data, called an archival information package (AIP), is separated and put into a storage location (e.g. hard drive or magnetic tape). Special metadata are stored in a "data management" database, to which administrative access can be given.

For a fully functional strategy for the long-term storage of electronic documents, it is absolutely necessary to compile the appropriate technical metadata. Unfortunately, no standard for a suitable metadata scheme specifically for long-term archiving has been developed for a long time. Therefore, Die Deutsche Bibliothek has introduced its own scheme, called Langzeitarchivierungsmetadaten für elektronische Ressourcen/Long-term Preservation Metadata for Electronic Resources (LMER)[6], derived from a model at the national library of New Zealand. The mostly automatic extraction of technical metadata is based on results from the JHOVE-project (JSTOR/Harvard Object Validation Environment) and kopal is increasingly becoming an active counterpart and contributor in the software development (Neubauer and Wollschläger, 2006).

To get an open, and enhanceable solution, various types of partners are taking part in the project. And because of differing motivations, the partners decided to keep the aspect of system maintenance separate from development. The partner responsible for the operation of the system (the computing center Gesellschaft für wissenschaftliche

Packaging

kopal
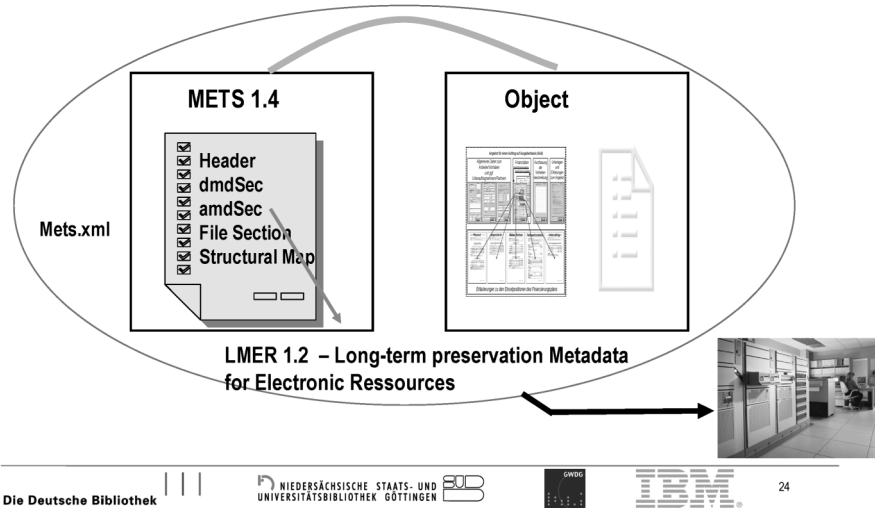*Data into the future*

Figure 1.
Universal object format

Datenverarbeitung Göttingen, GWDG) is hoping to gain experience with a well documented and scaled system, because it wants to attract further partners, who will use the benefits of the system for their own needs. On the one hand there is a neutral platform, where the developing partners Die Deutsche Bibliothek (DDB) and Goettingen State and University Library (SUB Göttingen) have to find common solutions for software and preservation procedures. On the other hand the system has to become capable of handling multiple users in a way that guarantees the independence of individual partners.

Within the project two of the project partners (DDB and SUB Göttingen) are putting digital material of all kinds into the long-term archive via batch processes. This ranges from digital documents in PDF, TIFF or TeX format to complex objects such as digital videos. After installing V2.0 of DIAS, a result of the first months of project-based software development including especially the realisation of universal object format capability, the project members are now starting the ingesting procedures in order to load multiple objects into the system. At the same time, some efforts have been made to establish a presentation system in DDB, which is based on a special caching area (here a server) that buffers used objects in an access area and delivers the objects rapidly according to user needs.

Regarding the software architecture, there is a separation between the core functionality of the archiving system and the environmental tools, which handle the homogenisation and the transfer of digital objects into the system. This task was taken on by DDB and SUB Göttingen using a cooperative and modularised concept, based on

JAVA-classes. The tools for building standardised submission information packages (SIPs) and for importing them into the system come with an open source licensing (GPL) method. The free software "kopal Library for Retrieval and In-gest" (koLibRI), with which archival objects can be created according to the UOF, will be available for public testing and analysing from March 2006[7]. For the presentation system the same principle applies: results must be independent from special dedicated solutions, based on well-defined interfaces, and open to other partners and systems (see Figure 2).

The development partner for the enhancement of the DIAS V.1.0 to the DIAS-Core (DIAS V.2.0) is IBM Germany GmbH. This will ensure a professional adoption of software components and provide stable long-term support. The separation of a core functionality (DIAS-Core) demands well-defined and freely available interfaces for future partners.

Another important goal of the project is the development of business models in the sense of organised cooperation and dedicated licensing solutions, in order to deliver multiple and flexible solutions for heterogeneous partners. Therefore, kopal has integrated various partners at different locations from the outset. In the future, this long-term archive for digital information will therefore provide other institutions with the opportunity to keep their data available on a long-term basis. Consequently, kopal ensures the possibility of academic, business and administrative use extending beyond libraries. On the one hand, there is the possibility of a client having its own "locker" in order to use the system with a secure storage space under its own administrative control. This solution is especially appropriate for small organisations or ones with a small amount of material to be archived. On the other hand, there is the possibility of later use of the kopal solution by installing the DIAS Core, which can be run together
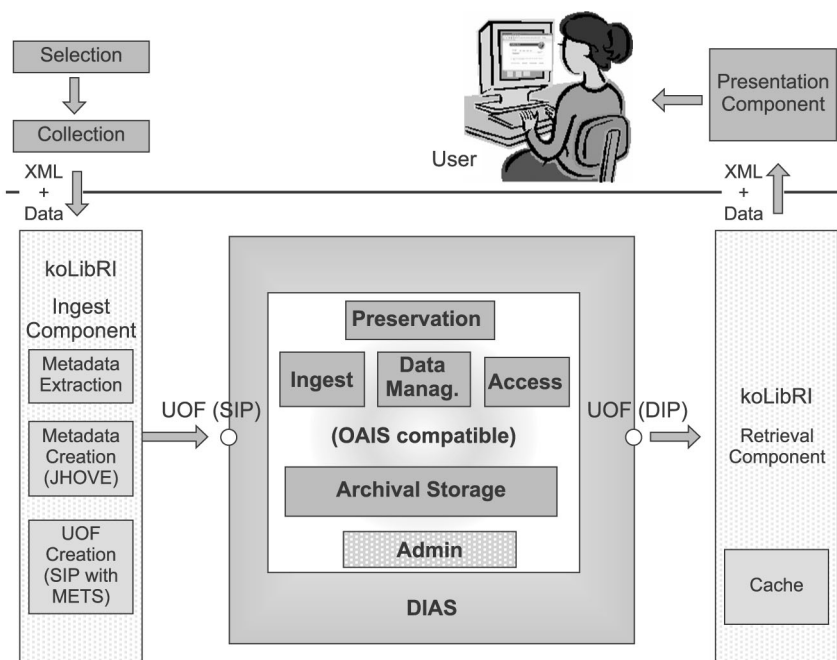
Figure 2.
Architectural overview

with kopal tools developed and supported by kopal. Additionally, as part of the project a working network/working group with the national library of The Netherlands was established in order to promote software innovations and strategic developments.

With the new architecture – sharing resources and spreading the use of the system – DIAS, in the form of the kopal solution, can become a central point in the worldwide search for solutions and strategies for preservation planning. The most important point in the project planning for the next 18 months is therefore the dedicated development of a detailed storage and service concept, the development of migration management tools and finally the finding, testing and systematic implementing of emulation strategies and tools.

Notes

1. www.ddb.de/
2. www.persistent-identifier.de/?lang = en
3. www.langzeitarchivierung.de/index.php?newlang = eng
4. http://www.langzeitarchivierung.de/modules.
   php?op = modload&name = PagEd&file = index&page_id = 18
5. http://kopal.langzeitarchivierung.de
6. www.ddb.de/standards/lmer/lmer.htmwww.ddb.de/standards/lmer/lmer.htm
7. http://kopal.langzeitarchivierung.de/index_koLibRI.php.de

Reference

Neubauer, M. and Wollschläger, T. (2006), "Maschinelle Gewinnung technischer Metadaten für die Langzeitarchivierung elektronischer Publikationen", *B.I.T. Online*, Vol. 1, pp. 37-40.

Corresponding author

Reinhard Altenhöner can be contacted at: reinhardaltenhoener@dbf.ddb.de

DINI institutional repository certification and beyond

Susanne Dobratz
Humboldt-University, Berlin, Germany, and
Frank Scholze
Stuttgart University Library, Stuttgart, Germany

Abstract

Purpose – The purpose of this article is to provide an overview on certification of institutional repositories as a means to support open access in Germany and a description of the DINI Certificate 2006 developed by DINI, the German Initiative for Networked Information.

Design/methodology/approach – The "DINI certificate for document and publication repositories" shows potential users and authors of digital documents that a certain level of quality in operating the repository is guaranteed and that this distinguishes it from common institutional web servers. The certificate can also be used as an instrument to support open access.

Findings – Repository certification will not be the main factor in achieving open access to academic information globally, but it can support the spread of institutional repositories and enhance visibility of the "institutional repository"-service.

Research limitations/implications – The DINI Certificate as a "soft" certificate aims towards interoperability of digital repositories, the coaching idea prevails. It does not provide an exhaustive auditing tool for trusted digital long-term preservation archives.

Practical implications – The "DINI certificate for document and publication repositories" pushed the development of institutional repositories in Germany according to certain organisational and technical standards and contributes to the interoperability amongst digital repositories worldwide.

Originality/value – This paper describes a unique approach that has been implemented in Germany and could be transferred to other countries and communities.

Keywords Digital storage, Archives management, Germany

Paper type Conceptual paper

Introduction: about DINI

DINI[1], the German Initiative for Networked Information (Deutsche Initiative für Netzwerkinformation) is a coalition formed by German Higher Education infrastructure and service institutions, such as libraries, computing centres and media centres, as well as by scientific learned societies. DINI itself is not a funding body, but it co-operates with German funding agencies like the German Research Foundation (DFG) or the German Ministry for Education and Research (BMBF).

DINI's primary objective is to create recommendations for standardised and interoperable information services and communication networks in and between universities. Therefore, the structuring and construction of networked digital publication services is one major task. DINI does this by evaluating and disseminating examples of good or best practise and by initiating and intensifying regional, national and international collaboration. International ideas, developments, and technologies, such as the Open Archives Initiative Protocol for Metadata

Harvesting (OAI-PMH), are observed and adapted for use in Germany. In this case the DINI Electronic Publishing Working Group formulated usage guidelines for institutional repositories in Germany (DINI Electronic Publishing Working Group, 2005).

DINI's focus is therefore the development and support of an information infrastructure. DINI is accomplishing this through operating a number of working groups[2].

The DINI Electronic Publishing Working Group is the most active group within DINI, having started its activities immediately after DINI was established in 2000. The members come from German universities and higher education institutions. Notable activities of the group were the organisation of workshops, such as the OAI Tutorials in Germany in 2003, the Workshop on the Implementation of Digital Repositories in Frankfurt in 2004 or the Open Access Symposium[3] in Göttingen in 2005. This event marked the beginning of open access activities at German universities. It was followed by the International Workshop on Institutional Repositories and Enhanced and Alternative Metrics of Publication Impact in Berlin, 2006[4], that focused on issues of visibility and the impact of scientific publications.

Certification in the DINI context

The "DINI certificate for document and publication repositories" (DINI Electronic Publishing Working Group, 2003) distinguishes the repository from common institutional web servers and assures potential users and authors of digital documents that a certain level of quality in repository operation and services is guaranteed.

The primary objective of the guidelines and criteria is:

- to improve interoperability and co-operation between German higher education institutions that run digital repositories; and
- to provide a tool for repository operators that could be used to raise the visibility, recognition, and importance of the digital repository within the university.

In addition, DINI sees its certificate as a tool for supporting the open access concept. It is regarded as a "soft certificate" (Dobratz and Schoger, 2005), focusing on the concept of coaching. Thus DINI defines certification slightly differently from others, who focus more on long-term preservation aspects (Dale, 2005; Ross and McHugh, 2005). One of the initiatives working on an auditing tool for certifying trusted digital repositories in the context of long term preservation is the Research Libraries Group/National Archives and Records Administration (RLG/NARA) task force on digital repository certification (see RLG/NARA Task Force on Digital Repository Certification, 2005; RLG Working Group on Digital Archive Attributes, 2002). Also the Network of Expertise in Long-Term Storage of Digital Resources (nestor)[5] for Germany (Dobratz et al., 2005), has established a working group on trusted repository certification, that is working on issuing a list of criteria for trusted digital long-term preservation repositories. The RLG/NARA audit checklist and the nestor certificate aim to document the trustworthiness of digital repositories.

Trustworthy digital repositories as defined by nestor can assure authors of all kinds of digital objects that their content is secured and preserved in a manner that ensures their authenticity and data integrity. It also provides a certain confidence for end users

that the information will be accessible over time and that the objects received from the digital repository are trustworthy in terms of the authenticity of the objects, the author, and publication time and place. For the institution itself and its co-operating partners, the certificate guarantees the reliability of the digital archiving services, which is a prerequisite for its integration into the overall mission of the institution and for collaboration on a national or international level.

DINI certificate for document and publication repositories

The DINI certificate aims at networking document and publication repositories by promoting the use of standards, interoperability and cooperation between German higher education institutions running digital repositories. By installing and running document and publication repositories, universities are able to offer and to archive scholarly publications that have been produced in-house, and make them available to a world-wide audience. This new service offered by the universities' information infrastructure units helps to disseminate the concept of electronic publishing as a new tool for academic work (Schirmbacher, 2005).

Since 1997, digital repositories have been developed in Germany with funding from the German Research Foundation (DFG) or the German Ministry of Education and Research (BMBF) (Scholze and Stephan, 2002). The reason for initiating a certificate for repositories was the result of a survey, which was conducted in 2003 by members of the DINI Electronic Publishing Group. A total of 47 German universities answered questions regarding the technology and standards used for their document servers. Summarising the results, a very disappointing picture was drawn concerning the use of standards for metadata, interfaces and cataloguing. In order to establish a German infrastructure for document repositories, DINI worked out criteria and guidelines; the "DINI certificate for document and publication repositories". With this certificate DINI provides a tool for repository operators that can be used to raise the visibility, recognition and importance of the digital repository within the university.

The certificate shows potential users and authors of digital documents that a certain level of quality in operating the repository is guaranteed, and that this distinguishes it from common institutional web servers. The DINI criteria are split into two sections. The first section specifies minimum standards and requirements that must be met by the document and publication repositories or their operators in order for them to be awarded the certificate. The recommendations are, as far as we can judge today, likely to become future requirements for the certificate. The auditing process is based on self-disclosure by the repositories and is conducted by two domain experts who are announced by the DINI office for each individual certification process.

The requirements and recommendations cover the following topics:

- visibility and server policy;
- author support;
- legal issues;
- authenticity and integrity;
- indexing (subject indexing, metadata, interfaces);
- visibility/impact/access statistics; and
- long-term availability.

A working group within DINI audits the criteria for the DINI certificate against international standards and developments and updates them accordingly. This continuous adaptation of the recommendations and requirements will ensure that the certificate retains its validity in an ever-changing environment.

For this reason the certificate is issued with a year-of-award stamp. The DINI office is responsible for awarding the DINI certificate – the document that acknowledges that the certified repository meets the minimum standards of a DINI certified document and publication repository. A small fee (€50-250) is charged for issuing the DINI certificate.

Practical experiences
So far, 17 university repositories have been awarded the DINI certificate[6], two repositories are currently being audited. It is often reported that the certification procedure has caused local authorities to reflect more deeply about the repository service itself and to start thinking about the repository's mission and philosophy. The certification recommendations and guidelines have proved to be a good way to bring the local repository up to a certain level of quality and to bring this to the attention of the institution's management.

Within the minimum requirements the following issues seem to be the most difficult, as has been reported by the operators of certified repositories:

  · introduction of a server policy, because the institution's governing body has to be involved;
  · provision of a visible service for authors; and
  · implementation of persistent identifiers, such as the URN:NBN schema[7] used in Germany and provided by Die Deutsche Bibliothek, the German national library (Schroeder, 2003).

Supporting open access through the DINI certificate 2006
The second issue of the certificate, the DINI certificate 2006 (DINI Electronic Publishing Working Group, 2006), stays abreast of changes and focuses on giving guidance for institutional repository operators to help them to support the open access concept and in particular to position their repository within the "green way" (Harnad, 2001), but it also gives recommendations for supporting the "golden way to open access" (Guédeon, 2004; Harnad, 2005).

In particular, it concentrates on the service an institutional repository, as an organisational and technological unit, can offer to support local academics in supplying pre- and post-prints of journal articles, that have already been published in scientific journals.

The following paragraphs describe the ideas for the new DINI certificate 2006. They are still subject to change within the ongoing editing process:

Visibility and server policy
In order to enhance the visibility of a document and publication repository, the range of services offered must be accessible via a single WWW-based entry point and a reference from the institution's home page to the repository's main entry point must exist. It is also recommended that the repository should be registered with an

aggregating service, such as the Institutional Archives Registry[8] or the Directory of Open Access Repositories – OpenDOAR[9].

There are also requirements on the server policy. It must state which standards are provided for publications in the repository with regard to content and functional and technical quality. An archiving guarantee for defined time spans, depending on the content and the functional as well as technical quality of the publications, has to be defined, as well as procedures for the operation of the repository. In particular, the services that the operator of the document repository offers to authors and editors must be published.

A DINI certified repository must define an open access policy containing a clear commitment to support the "green way" to open access.

Author support
It is vital that consultancy services and support for local authors (academics) should be offered via web pages, e-mail and telephone, as well as person-to-person support. Such support should be given for the entire publication process (including technical and legal areas).

In order to support the "green way" to open access the repository service must issue action guidelines for authors regarding secondary publications, provide self-upload for pre- and post-prints and, as a minimum, guide the authors with a link to the SHERPA/ROMEO list on publisher's policies[10].

It is recommended that a curriculum of courses on electronic publishing should be offered at least once per semester, as well as specialised courses on "structured writing" for authors. The utilisation of (electronic) help-desk systems and the provision of English language interfaces and descriptions may improve the quality and efficiency of author support.

The supply of style sheets or templates, of help texts that can be downloaded by authors (e.g. to produce PDF files), of references on how to use and cite documents as well as the provision of references to intellectual property rights and copyright is recommended.

Legal issues
The requirements for repositories are as follows. The operator of the repository must be permitted:

- to publish the uploaded document in the repository;
- to forward the document to an archiving institution; and
- to alter the documents technically to secure long-term availability.

In addition, exemption from liability must be formulated in a disclaimer, and the operator must openly demonstrate, that the documents are protected by intellectual property rights, copyright or licenses.

Furthermore it is recommended that rights and privileges should be listed in the metadata, as has been done at the University of Tübingen[11], and that authors should be offered a licensing tool during the upload process, as is done with Creative Commons[12] and additional licenses at the University of Tübingen with the local institutional repository TOBIAS-lib.

Authenticity and data integrity
In terms of ensuring a certain security, there are firstly requirements aiming to support the security of the repository. These include the documentation of the technical system (it is expected that the repository will be available 24 hours a day for seven days a week), the provision of a reliable back-up system to secure the repository, metadata and documents, and the existence of mechanisms that allow a technically controlled and verifiable acceptance of documents. In addition, the use of standard security mechanisms like SSL certification is recommended.

Second, the security of the documents has to meet a certain level. This means assigning persistent identifiers like uniform resource names (URN) or digital object identifiers (DOI) to the documents. It also means that a document with altered content must be treated as a new document, and therefore receive a new persistent identifier. Archiving the authors' uploaded files in their original format is also mandatory for a DINI certified repository. Recommendations for securing the documents themselves are the usage of an advanced digital signature technology, as provided by some vendors in Germany like Telesec[13], D-Trust[14] or Signtrust[15], according to the German Digital Signature Act (2005), or procedures to control the integrity of documents, such as hash algorithms.

Thirdly, in order to ensure the long-term availability of the documents, it is highly recommended that archiving file formats be rendered in order to export documents into long-term archiving facilities or institutions, such as kopal (Co-operative Development of a Long-Term Digital Information Archive)[16].

Indexing (subject indexing, metadata, interfaces)
Measures for enhancing the visibility of the documents and the servers include metadata, subject and formal indexing as well as interfaces for metadata exchange. As there are national classification schemas (e.g. the Regensburger Verbundklassifikation/RVK[17]) and normalised vocabulary for subject indexing terms (Schlagwortnormdatei/SWD[18]), DINI recommends, or even requires, the usage of these in order on the one hand to support a structured metadata exchange between the repositories, and, on the other hand, to enable libraries to treat their electronic publications like all other material held in their catalogues. By applying common library rules to the document and publication repositories, they are introduced into the usual library workflow and therefore receive more recognition by the library staff themselves. As digital publications are normally maintained within separate systems, using special repository software like E-Prints.org[19], D-Space[20] or OPUS[21], which, by default, are not automatically integrated into online public access catalogues (OPAC) or library network catalogues, it has to be actively and deliberately decided by the repository operators to use the appropriate features. E.g. OPUS 3.0 uses an adapted OAI-interface to exchange enriched XMETADISS data with the union catalogue of the South-West German Library Network (SWB).

For subject indexing, DINI requires:

- the availability of a defined policy which should be made known to authors;
- the usage of verbal indexing with keywords or classificatory indexing; and
- the use of the Dewey decimal classification (DDC) according to usage in the German National Bibliography[22] as the general classificatory indexing system for documents (see also DINI Electronic Publishing Working Group, 2005).

Recommendations covering subject indexing are: the use of at least one additional standardised system of verbal or classificatory indexing (general or subject specific), e.g. Schlagwortnormdatei (SWD), Library of Congress subject headings, ACM computing classification system (CCS), mathematical subject classification (MSC), or physics and astronomy classification scheme (PACS). In order to make German documents internationally accessible and searchable, the provision of keywords in English and abstracts in German and English is considered important.

The following issues regarding the use and availability of metadata are required: firstly, the metadata must be available for free, and secondly as a minimum, the metadata should be encoded as Unqualified Dublin Core (ISO, 2003) according to OAI-PMH (DINI Electronic Publishing Working Group, 2005; Lagoze et al., 2002).

As a recommendation for metadata, DINI suggests the use of Qualified Dublin Core, ONIX (Online Information eXchange)[23], and the enhancement of metadata with technical and archival metadata, using standards like metadata encoding and transmission standard (METS), long-term preservation metadata for electronic resources (LMER) or preservation metadata: implementation strategies (PREMIS) (see Die Deutsche Bibliothek, 2006; METS Editorial Board, 2005; PREMIS Working Group, 2005). Also recommended is the enhancement of metadata with special service metadata like print-on-demand data, e.g. PROPRINT (see Mittler and Schulz, 2004).

It is also highly recommended that metadata and documents should be able to be exported to long-term archiving institutions or repositories and that metadata should be able to be exported to bibliographical databases (e. g. bibliographical management system, library networks, OPACs).

In order to enhance the availability of the documents via services like Google or SCIRUS[24], or via scholarly search engines like Bielefeld Academic Search Engine (BASE)[25], DINI suggests the provision and promotion of special link lists for indexing by robots and commercial search engines.

Moreover, a DINI certified repository provides a worldwide web end-user interface and an OAI-PMH 2.0 interface in accordance with the DINI recommendations on OAI-PMH use in Germany (DINI Electronic Publishing Working Group, 2005).

It also uses extended metadata schemas with OAI 2.0 to allow for the exchange of complex metadata schemata and additional web service interfaces (e. g. Simple Object Access Protocol/SOAP) on a voluntary basis. The use of a Z39.50 interface is also recommended as it gives the chance of integrating the repository with library networks, which mainly operate on this particular protocol.

Visibility – impact – access statistics
As there are intensive discussions on introducing a way of measuring the impact and use of electronic documents in institutional repositories, DINI sees the necessity of helping repository operators to introduce such a technology, once it has been approved. Possible technologies and approaches are discussed, amongst others, by Bollen et al. (2005), Brody (2006), Hardy et al. (2005), Harnad et al. (2004), and Lawrence (2001).

A prerequisite for providing reliable usage statistics are normalised webserver log files. A recommendation that is likely to become a standard in the next edition of the DINI certificate is the use of the counting online usage of networked electronic

resources (COUNTER) model for processing access logs from electronic journals or e-books. The COUNTER model cannot, at present, be assigned to all institutional repositories and the material available there in the same way, because issues like web crawler access have not yet been resolved. The DINI Electronic Publishing Working Group dedicates itself to supporting the development of such a model in co-operation with COUNTER and other interested parties like project (Interoperable Repository Statistics (IRS) (see COUNTER, 2005, 2006). For the time being DINI recommends the use of existing lists to filter non-human access[26].

The requirements demanded by DINI are as follows: each individual repository must (within the limits of the law) log statistical data on access to both the repository and to individual publications. The webserver logs have to be anonymised before they are stored for long periods of time.

In order to make usage visible to users and authors, it is necessary to link individual publications with their access statistics as dynamic metadata.

It is essential that access statistics be accompanied by documentation explaining which criteria were used to collect the basic data and how it was processed. Furthermore, it has to be indicated on the web pages that access figures that are not collected and processed in a universal and standardised way cannot be used to compare different repositories. Access figures, as they are published by the repositories today, can only be used to compare the access figures for documents in the same repository.

Long-term availability

Long-term preservation has become an important issue within the electronic publishing process. However, it is neither necessary nor possible, from an economic point of view, to turn every institutional repository, or every document and publication repository, into a trusted digital long-term preservation archive. This highlights the need for global co-ordination of institutional repositories on the one hand, and trusted digital archives on the other hand. This co-ordination demands the preparation of digital documents and their metadata as Submission Information Packages according to the open archival information system (OAIS) (CCSDS, 2002).

DINI supports this concept and requires the persistent linking of metadata and documents (e.g. via a persistent identifier, or the storage of the metadata and document in one single container), the inclusion in the repository's policy that the minimum availability of a document be no less than five years, and the provision of archive copies that are free of digital rights management (DRM) measures, which prevent the use of long-term preservation strategies (migration, emulation).

Recommendations for the support of the long-term availability of digital documents are: to take steps to secure long-term availability, where necessary through co-operation with an archiving institution, and to support and promote the use of open file formats for long-term preservation (e.g. PDF/a instead of PDF).

The production of technical metadata for long-term preservation (e.g. using tools like the JSTOR/Harvard Object Validation Environment/JHOVE) is regarded as important preparation for the later execution of preservation strategies. This should be connected with the unique identification of the used file format in the metadata, with reference to public file format registries.

Conclusion

Repository certification by itself can certainly not accomplish the open access idea within universities and research institutions. As long as scholars and scientists continue to fail to adopt the open access concept, universities will not succeed in implementing open access, even if they were to offer perfectly functioning institutional repositories from a technological point-of-view.

It is therefore essential to promote the scholarly relevance of the institutional repository in order to reach a critical mass of documents. The most convincing factor in encouraging authors to follow open access is acknowledgement by colleagues. This is achieved by counting and measuring usage and citation data of publications. It has already been shown that for most disciplines the impact factor is increased if articles are available through open access (Harnad *et al.*, 2004; Lawrence, 2001). These findings have to be publicised among academics. Another potential approach to advance on the "green" as well as the "golden road" to open access is to establish impact factors that take into account open access publications and to give them the same position as existing metrics like the ISI impact factor (IF)[27].

There are interesting suggestions for alternative metrics of impact and usage as has been shown in the DFG/DINI workshop on impact measures[4] in February 2006 (see also Ball, 2006). DINI will discuss these suggestions taking into account the basic data that has to be collected as well as the methods and algorithms applied to that data. DINI will implement selected approaches in a German "testbed" consisting of DINI certified servers.

In summary, repository certification will not be the main factor in achieving open access to academic information globally, but it can facilitate the spread of institutional repositories and enhance visibility of the "institutional repository" service.

It is obvious from DINI's perspective, that universities and research institutions need guidance in installing institutional repositories. Institutions need to reach their academic staff, as they are the authors, and promote the quality and add-on functionalities that can be obtained by using an institutional repository instead of a personal home page.

DINI, in co-operation with other German players such as the German Rectors' Conference (HRK), the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK), the German Research Foundation (DFG) and the German Ministry of Education and Research (BMBF), will engage in providing information space, support for post- and pre-print issues, and guidance on Open Access policies and copyright issues for German Higher Education and research institutions. DINI will incorporate international developments and ideas, and will co-operate with projects like OpenDOAR, Securing a Hybrid Environment for Research Preservation and Access (SHERPA) and Digital Repository Infrastructure Vision for European Research (DRIVER) (Lossau, 2006).

Nevertheless, standardisation and interoperability are still badly needed in order to build service layers on top of institutional repositories. The provision of a self-upload tool or self-cataloguing by authors are first steps, access statistics are a key issue for visibility and a step towards alternative impact metrics. DINI's expertise lies in the support and marketing of interoperability technologies for information infrastructures. For this reason, DINI will continue its work on repository certification and will enrich this approach with active partners world-wide in order to launch additional

infrastructural activities to promote open access. DINI plans to form a core group to develop an open access advocacy package, to support the "green way" by integrating more information about German publishers' open access policies into SHERPA, to offer advice for authors, librarians, and especially for university management and funding bodies.

On a technical level, DINI plans to enhance interoperability between German open access repositories, and to develop an infrastructure to measure impact factors, which takes into account repository content as well as journal content based on internationally agreed standards, technologies and methods.

Notes

1. www.dini.de
2. www.dini.de/dini/arbeitsgruppe/arbeitsgruppen.php
3. www.dini.de/veranstaltung/workshop/goettingen_2005-05-23/
4. www.dini.de/veranstaltung/workshop/oaimpact/
5. www.longtermpreservation.de
6. www.dini.de/dini/zertifikat/zertifiziert.php (last visited on 14.03.2006)
7. www.persistent-identifier.de
8. http://archives.eprints.org/
9. www.opendoar.org
10. www.sherpa.ac.uk/romeo.php
11. http://w210.ub.uni-tuebingen.de/dbt/uni/licenses.php
12. http://creativecommons.org/
13. www.telesec.de/
14. www.d-trust.net/
15. www.signtrust.de
16. www.kopal.langzeitarchivierung.de/
17. www.bibliothek.uni-regensburg.de/Systematik/systemat.html
18. www.ddb.de/standardisierung/normdateien/swd.htm
19. www.eprints.org
20. www.dspace.org
21. http://elib.uni-stuttgart.de/opus/doku/opus_sw.php
22. www.ddc-deutsch.de/
23. See description at www.editeur.org/
24. www.scirus.com
25. http://base.ub.uni-bielefeld.de/index_english.html
26. See, e.g. www.robotstxt.org/wc/active/all.txt
27. Via Thompson Scientific, Web of Knowledge http://scientific.thomson.com/webofknowledge/aboutwok.html

References

Ball, P. (2006), "Prestige is factored into journal ratings: 'Y-factor' measures quality as well as quantity of citations", *Nature*, Vol. 439 No. 7078, pp. 770-1.

Bollen, J., Van de Sompel, H.S.J.A. and Luce, R. (2005), "Toward alternative metrics of journal impact: a comparison of download and citation data", *Information Processing and Management*, Vol. 41 No. 6, pp. 1419-40.

Brody, T. (2006), "Citation impact of open access articles vs. articles available only through subscription ('toll-access')", ongoing study, available at: www.citebase.org/isi_study/ (accessed 27 April 2006).

Consultative Committee for Space Data Systems (CCSDS) (2002), "Reference Model for an Open Archival Information System (OAIS)", *Blue Book*, available at: http://public.ccsds.org/ publications/archive/650x0b1.pdf (accessed 24 April 2006).

Counting Online Usage of Networked Electronic Resources (COUNTER) (2005), "COUNTER code of practice for journals and databases: release 2", available at: www.projectcounter. org/code_practice.html (accessed 27 April 2006).

Counting Online Usage of Networked Electronic Resources (COUNTER) (2006), "COUNTER code of practice for books and reference works release 1", available at: www. projectcounter.org/code_practice.html (accessed 27 April 2006).

Dale, R.L. (2005), "Making certification real: developing methodology for evaluating repository trustworthiness", *RLG DigiNews*, Vol. 9 No. 5, available at: www.rlg.org/en/page. php?Page_ID = 20793#article2 (accessed 26 April 2006).

Die Deutsche Bibliothek (2006), "LMER: long-term preservation metadata for electronic resources", available at: www.ddb.de/standards/pdf/lmer12_e.pdf (accessed 26 April 2006).

DINI Electronic Publishing Working Group (2003), "DINI certificate for document and publication repositories", available at: www.dini.de/documents/Zertifikat-en.pdf (accessed 27 April 2006).

DINI Electronic Publishing Working Group (2005), "Electronic publishing in higher education: how to design OAI interfaces – recommendations", 2nd edition, available at: http://edoc. hu-berlin.de/series/dini-schriften/2005-2-de2/PDF/2-de2.pdf (accessed 27 April 2006).

DINI Electronic Publishing Working Group (2006), "DINI certificate for document and publication repositories", in press.

Dobratz, S. and Schoger, A. (2005), "Digital repository certification: a report from Germany", *RLG DigiNews*, Vol. 9 No. 5, available at: www.rlg.org/en/page.php?Page_ID = 20793#article3 (accessed 26 April 2006).

Dobratz, S., Neuroth, H., Schoger, A. and Strathmann, S. (2005), "nestor – Entwicklungsstand des Kompetenznetzwerks zur Langzeitarchivierung digitaler Ressourcen in Deutschland", *ZfBB – Zeitschrift für Bibliothekswesen und Bibliographie*, Vol. 52 Nos 3/4, pp. 151-62.

Guédeon, J.-C. (2004), "The 'green' and 'gold' roads to open access: the case for mixing and matching", *Serials Review*, Vol. 30 No. 4.

Hardy, R., Oppenheim, C., Brody, T. and Hitchcock, S. (2005), "Open access citation information", available at: http://eprints.ecs.soton.ac.uk/11536/ (accessed 27 April 2006).

Harnad, S. (2001), "The self-archiving initiative", *Nature*, Vol. 410 No. 6832, pp. 1024-5.

Harnad, S. (2005), "Fast-forward on the green road to open access: the case against mixing up green and gold", *Ariadne*, No. 42, available at: www.ariadne.ac.uk/issue42/harnad/ (accessed 26 April 2006).

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Stamerjohanns, H. and Hilf, E. (2004), "The access/impact problem and the green and gold roads to open access", *Serials Review*, Vol. 30 No. 4, pp. 310-4.

International Standardisation Organisation (ISO) (2003), "Information and documentation – the Dublin Core metadata element set ANSI/NISO Z39.85-20012 or ISO 15836:2003", available at: www.niso.org/international/SC4/n515.pdf (accessed 27 April 2006).

Lagoze, C., Van de Sompel, H. and Nelson, M.W.S. (2002), "The open archives protocol for metadata harvesting", available at: www.openarchives.org/OAI/openarchivesprotocol.html (accessed 27 April 2006).

Lawrence, S. (2001), "Free online availability substantially increases a paper's impact", *Nature*, Vol. 411 No. 6837, p. 521.

Lossau, N. (2006), "Digital repository infrastructure vision for European research", paper presented at the International Workshop on Institutional Repositories and Enhanced and Alternative Metrics of Publication Impact, Humboldt University, Berlin, 20-21 February 2006, available at: www.dini.de/veranstaltung/workshop/oaimpact/presentations/lossau.ppt (accessed 27 April 2006).

METS Editorial Board (2005), "METS: metadata encoding and transmission standard: schema and documentation", available at: www.loc.gov/standards/mets/ (accessed 27 April 2006).

Mittler, E. and Schulz, M. (2004), "ProPrint world-wide print-on-demand services for study and research", *Library Hi Tech*, Vol. 22 No. 2, pp. 227-30.

PREMIS Working Group (2005), "Data dictionary for preservation metadata", available at: www.oclc.org/research/projects/pmwg/premis-final.pdf (accessed 27 April 2006).

RLG/NARA Task Force on Digital Repository Certification (2005), *Audit Checklist for Certifying Digital Repositories*, RLG/NARA Task Force on Digital Repository Certification, Mountain View, CA.

RLG Working Group on Digital Archive Attributes (2002), *Trusted Digital Repositories: Attributes and Responsibilities*, RLG, OCLC, Mountain View, CA, available at: www.rlg.org/legacy/longterm/repositories.pdf (accessed 27 April 2006).

Ross, S. and McHugh, A. (2005), "Audit and certification of digital repositories: creating a mandate for the Digital Curation Centre (DCC)", *RLG DigiNews*, Vol. 9 No. 5, available at: www.rlg.org/en/page.php?Page_ID = 20793#article3 (accessed 26 April 2006).

Schirmbacher, P. (2005), "Die neue Kultur des elektronischen Publizierens", in Nielsen, E.K., Saur, K.G. and Ceynowa, K. (Eds), *Die innovative Bibliothek*, K.G. Saur, Munich, pp. 107-20.

Scholze, F. and Stephan, W. (2002), "Electronic publishing", in Leonhard, J.-F. (Ed.), *Handbücher zur Sprach- und Kommunikationswissenschaft*, Vol. 15, De Gruyter, Berlin, pp. 2634-48.

Schroeder, K. (2003), "Persistent Identifiers im Kontext von Metadaten", *ZfBB – Zeitschrift für Bibliothekswesen und Bibliografie*, Vol. 50 No. 4, pp. 205-9.

Corresponding author
Susanne Dobratz can be contacted at: dobratz@cms.hu-berlin.de

ProSeBiCA: development of new library services by means of conjoint analysis

Antonia Hermelbracht

Department of Business Administration and Economics, University of Bielefeld, Bielefeld, Germany, and

Bettina Koeper

Bielefeld University Library, Bielefeld, Germany

Abstract

Purpose – The purpose of this article is to provide a description of the project ProSeBiCA, carried out in order to adapt the marketing research tool conjoint analysis for the development of future library services. The paper describes the methodical approach, and provides an overview of the results gained by several user surveys.

Design/methodology/approach – The methodical approach includes the use of adaptive conjoint analysis (ACA) and choice-based conjoint analysis (CBC).

Findings – Conjoint analysis is an appropriate tool for the strategic development of library services. It provides detailed results by identifying users' preferences towards concrete services, and allows to deduce general statements about future trends for library services.

Research limitations/implications – Future plans include cooperation with the Sheridan Libraries at Johns Hopkins University (Baltimore, Maryland), the Association of Research Libraries (ARL), and other interested partners in order to adapt the ProSeBiCA framework for US libraries and to integrate techniques used by the partners (CAPM, LibQUAL $+^{TM}$). The extension of conjoint analysis to the evaluation of services for a fee will be an additional area of research.

Practical implications – Service orientation in libraries can be based on a systematic approach. Depending on local circumstances, interested libraries can improve their service orientation in different ways, either by discussing the results of the analysis made in this project or by using the Conjoint Analysis questionnaire for a survey in their institutions.

Originality/value – The paper describes the current status and provides the latest results of ProSeBiCA. It is of interest to library managers and marketing researchers.

Keywords Information services, Market research, Surveys, Academic libraries, Germany

Paper type Research paper

In the context of changing educational environments, current discussions about the strategic development of German academic libraries clearly show the need for a basic change in the way they see themselves. They need to evolve from mere academic institutions into service providers that actively design and offer services that fit users' needs and preferences. More and more customer take-up of library services is becoming a new quality standard that will have a significant effect on the status of libraries within universities, particularly in times of small budgets and growing autonomy in academia. In order to achieve this take-up, libraries need to establish and improve their customer relations, and above all they have to develop a profound knowledge of their users and their needs. But how can we get this knowledge in a

systematic way? How can we get a clear idea of user preferences in order to shape future library services?

The project ProSeBiCA, funded by the German Research Foundation (DFG) and carried out jointly by the Chair of Marketing at Bielefeld University and Bielefeld University Library, tries to answer these questions. The name "ProSeBiCA" is an acronym of the German project title that can be translated as "Prospective control of the services of academic libraries by means of conjoint analysis". It suggests the idea behind the project, which is to adapt conjoint analysis to create an appropriate tool for library managers. Now, what exactly is meant by conjoint analysis?

Conjoint analysis: a short definition

Conjoint analysis is a well-established marketing research method that aims to examine customer preferences empirically. It also aims to simulate the users' potential decisions on the use of products or services. In other words: conjoint analysis offers a mechanism for measuring user take-up, in particular, take-up of new products and services which have yet to be developed. Conjoint analysis does not formulate the problem in terms such as: "how does a customer value a certain service in its current form?", but "what service or what level of service will yield the greatest value to the customer?" and, in a second step, "which of these options should be incorporated into the future range of services?" Hence conjoint analysis is a proactive method that aims to give a picture of the future shape of services. It is widely used in commercial applications, mostly in the fields of pricing policy and development of new products. In contrast, it is not common in the public sector, with the exception of public health services.

First of all, there are a few technical terms connected with conjoint analysis that should be introduced. Conjoint analysis always distinguishes between "attributes" and "levels". Products or services (e.g. library news bulletins) are "attributes", and their potential forms (e.g. library news bulletins provided in print form on flyers, or in electronic form via e-mail) represent the "levels". Customers have to choose between these possible levels, thus expressing their preferences. Combining levels of different attributes allows service offerings to be defined. They can then be compared according to their value to customers. However, conjoint analysis is not only able to reveal the preferences of the customer-base as a single entity. It also allows group specific analysis. The "a-priori segmentation" looks at the preferences of respondents who have been divided into segments defined by demographic criteria (e.g. status, such as student, academic staff or administrative staff). By interpreting the results of a conjoint analysis survey it is also possible to group the respondents according to the "part-worth utilities" (output of the conjoint analysis representing the particular values of the single attribute levels) they ascribe to a product, this is the so-called "benefit segmentation".

Aims and procedures of the ProSeBiCA project

The aim of the ProSeBiCA project is to adapt conjoint analysis as a tool for the strategic development of academic libraries. This means the developing of an analysis and simulation framework that covers the whole range of library services. This aim is twofold. Firstly, a framework has to be developed for Bielefeld University, because a methodical adaptation such as is intended by this project clearly needs a test environment. Secondly, ProSeBiCA will provide a general analysis framework and some guidelines for the application of Conjoint Analysis in other academic libraries. In

this spirit, the project is to give an insight into what users may expect from their library and to identify future trends that are relevant for the development of library services.

A project like this one, with the intention of adapting a complex marketing research method for library services, clearly depends on a practical background as well as on academic management input. In this way it forms a perfect basis for the cooperation of a university library and faculty. Hence, ProSeBiCA is a cooperative project carried out jointly by Bielefeld University Library and the chair of marketing at the local department of economics and business administration. The project passed several milestones, the first one being to generate ideas. This step was preparation for the adaptive conjoint analysis survey (ACA) at Bielefeld University in Autumn 2004 as well as for the choice-based conjoint analysis survey (CBC) in Bielefeld from May to June 2005. The next milestone, the combined ACA and CBC survey at Brandenburg Technical University Cottbus, ran from November 2005 to January 2006. The latter survey covered the same questions as the previous two studies in Bielefeld, and the evaluation and interpretation of the results is currently (as of February 2006) still ongoing. This milestone, requiring cooperation with the Information, Communication and Media Center (IKMZ), as well as with the chair of marketing and innovation management at Cottbus University, is of great importance, as it is to ensure the transferability of the approach developed at Bielefeld. This finally leads on to the last milestone, which includes general statements and results, as well as the general framework and the guidelines already mentioned.

Alongside ProSeBiCA, there was an information exchange with the Sheridan Libraries at Johns Hopkins University, who applied a similar methodical approach in their project Comprehensive Access to Printed Materials (CAPM). Previous considerations about possible connections between CAPM and LibQUAL + ^{TM}, a methodology developed by the Association of Research Libraries (ARL) and Texas A&M University Libraries (Choudhury *et al.*, 2002; Heath *et al.*, 2003), also led to further contacts with ARL and other interested partners. As a result of these contacts a working plan was established to adapt the ProSeBiCA framework for US libraries. Taking a step forward, a plan was also established to integrate the techniques offered by these three methodical approaches (CAPM – LibQUAL + ^{TM} – ProSeBiCA) with the aim of developing a homogenous tool and portal as a support system for library management decision-making processes.

Empirical studies at the Universities of Bielefeld and Cottbus
Idea generation step
In order to cover all relevant, possible and desirable future library services, a variety of activities were started during Spring 2004. Firstly, a comprehensive content analysis of secondary data sources was carried out. These included, in particular, articles in library journals, political papers as well as websites of a broad range of libraries all over the world. Subsequently, the ideas and opinions of 1,349 academic library users at four different German universities concerning the services offered by academic libraries were collected by means of an empirical survey. In addition, new service ideas were generated by academics and librarians in a total of five structured brainstorming workshops. Furthermore, library usage data from Bielefeld University Library was analyzed by means of an artificial neural network approach in order to identify usage

patterns that support the qualitative new service creation process (cf. Decker and Hermelbracht, 2005). Altogether it was possible to identify some controversial strategic themes and – from an operative perspective – more than 250 concrete new ideas for services.

From the strategic perspective, it was possible to identify areas such as the degree of digitization at the library, the level of support for users by librarians, or the amount of services available for a fee. The six most relevant areas were integrated in two choice-based conjoint analysis surveys (CBC) at Bielefeld University and Cottbus University. This selection was also based on the results of the operative perspective (see below).

From the operative perspective, there was an abundance of ideas about many different areas. A detailed list of these concrete ideas for new services is given in Hermelbracht and Senst (2005). A database with an up-to-date pool of ideas is available from the ProSeBiCA-project homepage[1]. Two short examples serve to illustrate the spectrum of ideas. An example of a more technical service could be a video-mediated book-view, where users can read and scroll through a book online via a high-resolution video. Another innovation especially for new users could be the starter pack, including a virtual short introduction to using the library for freshers as well as a glossary for foreign students with translations of important library terms into several languages. In the end, the challenge of structuring and handling this variety of ideas appropriately was met by allocating the service ideas to the following areas: provision of media, learning and working environment, communication, and additional services. These areas are part of an integrated structure, into which any current or potential service of a university library can be clearly slotted. Provision of media, the learning and working environment and communication are part of the core business of a university library. Each of the four areas was evaluated separately by an adaptive conjoint analysis (ACA) that first took place in an empirical survey at Bielefeld University and then in a second one at Cottbus University.

Questionnaire
For the two surveys in Bielefeld (ACA and CBC), as well as the survey in Cottbus, we used the two worldwide most popular online conjoint software packages offered by Sawtooth Software. In applying the ACA it was possible to use its advantages, such as its handling of a large number of attributes and levels, dynamic question creation, and calibration of utilities for market simulations. The adaptive conjoint analysis contains compositional as well as decompositional parts and gathers the different parts of information using four kinds of questions. The choice-based conjoint analysis takes the choice-behavior of the consumers into account, and reflects new research developments in the area of conjoint analysis. CBC is more appropriate for scenarios with fewer attributes, as in the case of strategic perspectives. The respondent can compare different complete profiles and choose one of them. A description of these conjoint techniques in the context of the current project is given in Decker and Hermelbracht (2006). A visualization of the ACA questions can be seen in Hermelbracht and Senst (2006). Due to the very large amount of ACA attributes (42) and levels (118), they cannot be presented here. For more information, please refer to the homepage of the project.

The CBC questionnaire used one main type of question to discover the preferences of library users (see Figure 1). This study considered six attributes with two or three levels. The attributes were:

- innovation strategy;
- level of support;
- degree of digitization;
- degree of specialization;
- add-on services with costs, and
- presentation of services.

The three levels of the innovation strategy were reactive (demand-oriented innovations), selective (innovation in selected areas), and progressive (strict innovation-orientation). The level of support can vary from unsupported/independent working, via assisted working, to delegating tasks to librarians. In order to evaluate the preferred degree of digitization, the scenario of a conventional library with digital services was differentiated from that of a completely digitized library existing only in cyberspace. The level of specialization of regularly offered services which should be striven for in the future, can be decided from the following three options: no emphasized service specialization, service specialization in selected fields, or a focus on a few top performance services. The fifth attribute considers the role of chargeable secondary services, which can be provided to a small, medium or large extent. Lastly, the services can be presented in two opposing ways: either in a pragmatic and functional way, or in an entertaining and animating way. Which of these options were preferred in both surveys will be analyzed next in a short presentation of the survey results.

Implementation and general results
The first ACA survey at Bielefeld University (October to December 2004) collected 2,120 answers over an eight-week period. The second CBC survey at Bielefeld University took place from May to June 2005 and gathered 1,672 responses. Lastly, the integrated ACA and CBC survey at the University of Cottbus (November 2005 to January 2006) collected the opinions of 1,127 persons. Table I shows the distribution of respondents according to the five different questionnaire parts. Almost 5,000 responses were gathered during the three empirical surveys.

In order to obtain such large samples the questionnaire was placed on each library's homepage (see also Hermelbracht and Senst, 2006). Additionally some students helped to motivate visitors to the library and students on campus to take part in the survey, and also made phone calls to academics and to administrative staff explaining the purpose of the questionnaire. In addition, extrinsic motivation was increased by offering attractive prizes (e.g. iPod, gym memberships, theatre tickets, music CDs), which could be won by completing the questionnaire. All surveys covered the different demographic groups very well. The different user groups were represented as follows:

- in the ACA survey in Bielefeld, there were 1,685 students (equaling approximately 9 percent of the undergraduates, graduates, and postgraduates matriculated at Bielefeld University), 174 academic staff (equaling
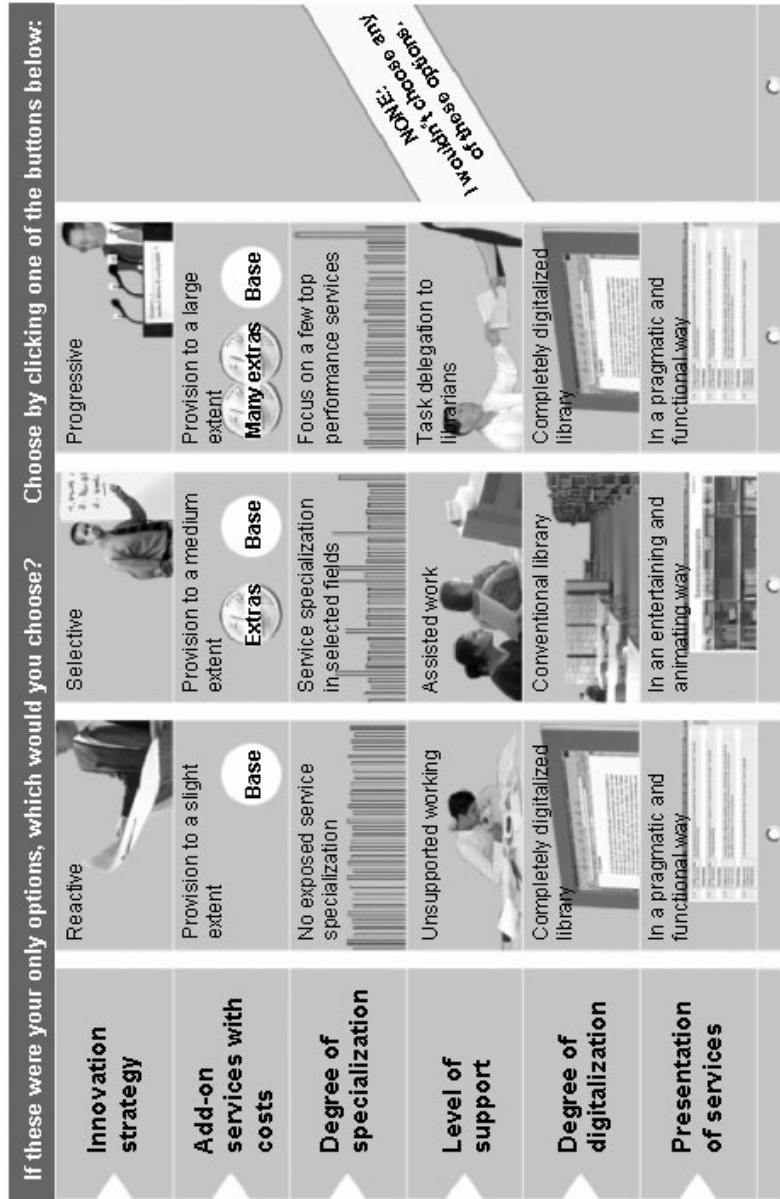
Figure 1.
Sample CBC choice task

approximately 22 percent of the professors and academic assistants), and 261 others (e.g. external users);
• in the CBC survey in Bielefeld, there were 1,329 students, 118 academics and 225 others; and
• in the integrated ACA/CBC survey in Cottbus, there were 843 students, 141 academics, as well as 143 others.

More details about the demographics can be found in the project documentation[1].

ACA results

Different aspects were investigated through the analysis of the ACA data. The most important ones considered the analysis of preferences on an aggregate level, on an a-priori user-groups level, and on an a-posteriori benefit-segments level. In the aggregate level of analysis it was discovered that many current services (such as the conventional library catalogue search, self-administration of the loan account and the online post box) are evaluated favorably, (for more details see Hermelbracht and Koeper, 2006). Also many improvements to current services are regarded as providing a high utility (for example, an increase in printed media stock or an improved guidance system). Finally, there is a demand for many innovative ideas, such as an open-air area or starter packs. On the other hand, there are conventional features that are less desirable, such as conventional interior design and a lack of online-publishing possibilities. There is also little demand for some new service ideas, such as a radio-archive or drive-thru book lending. It can be concluded that there are differences between the evaluation of current services and new innovative ideas on the aggregate level. Altogether it was possible to identify acceptable as well as controversial innovations and provide recommendations for the future development of Bielefeld and Cottbus University Libraries. A more detailed analysis also revealed similar discrepancies in the preferences of different user groups.

The a-priori segments (such as academics, students, and others) have some preferences in common. Clear preference structures could be identified in the same way as with the aggregate analysis. Academics are more skeptical towards more "extravagant" innovations than the students and others. They also make the strongest differentiation between the different attribute levels.

Based on the part-worth utilities of the respondents a cluster analysis was carried out. As a result it was possible to identify some benefit-segments. These vary between the different service areas as well as between the Universities of Bielefeld and Cottbus. But some response patterns were seen to recur in the different sub areas. Segments

	ACA Bielefeld 13.10.04-09.12.04	CBC Bielefeld 04.05.05-30.06.05	ACA/CBC Cottbus 03.11.05-04.01.06	Total
ACA: Provision of media	578	0	196	774
ACA: Learning environment	540	0	201	741
ACA: Communication	455	0	196	651
ACA: Additional services	547	0	184	731
CBC: Strategic development	0	1,672	350	2,022
Total	2,120	1,672	1,127	4,919

Table I.
Distribution of the responses to the three empirical surveys

with highly-involved users oriented towards innovation are typical for almost every consideration. Also, some traditional users who prefer many current services are represented in each sub area. In some of the areas, it was possible to identify groups of determined users, who differentiate very strongly between the single items, and less-interested users, who are unlikely to make great use of the services.

Comparison of the part-worth-utilities in Bielefeld and Cottbus reveals some interesting correspondences. The operative evaluations can be separated into two parts. Half of the results in Bielefeld and Cottbus are similar, and the other half contradict each other. Therefore it will be difficult to deduce general recommendations on concrete services for all libraries in Germany. Moreover, each library has its own specific situation, which should be investigated separately. It is interesting that the results become more similar if universal services such as different communication activities are considered. Then again, if more specific services such as from the learning and working environment area are considered, the differences increase.

CBC results
This impression can be confirmed by the results of the strategic CBC surveys. Here the results in both cities are very similar. The middle attribute levels are always considered as offering the highest utility, with the consequence that, for example, the users decided in favor of selective innovation strategy, assisted working (but no task delegation to librarians), and also specialization in selected areas. Here it is very important to stress the significance the users attribute to the conventional library as working environment with some digital media stock and services. This corresponds to the results of the ACA survey. The following outcomes also confirm the first survey results. The respondents prefer the pragmatic and functional to the entertaining and stimulating presentation of services – digital as well as those physically on site. This corresponds to the valuation of some "extras" included in the ACA survey. Finally it can be concluded that the respondents are willing to pay for some additional services, as long as these do not take precedence. A detailed consideration of the a-priori and benefit-segments, as well as extensive presentations of the ACA and CBC results, are contained in the documentation of the project and are available via the project homepage.

Conclusions
What can be said about the general outcomes of the ProSeBiCA project? It considers a very interesting and up-to-date theme. It was possible to collect very large samples and many respondents communicated their interest in this topic.

In all surveys it was possible to identify the high importance of the hybrid library, as well as of many on-site services. Altogether the significance of the core business, and especially of the provision of media, is very high. As to concrete innovations, we can clearly differentiate between those for which there is a strong demand, and others that obviously do not add any useful functionality. There is a willingness to accept new service ideas, and there are strong preferences, especially for services that facilitate immediate and effortless access to information. Alongside the benefit-segments, it was possible to identify some interesting differences between the a-priori segments, considering extras and facilities for the users.

Altogether, it can be concluded that abstract results are comparable and probably valid for other university libraries in Germany. The evaluation of concrete services depends strongly on the particular situation of the library under consideration. Libraries interested in marketing research can profit from the results of this project as a basis for future planning and are well advised to deepen their knowledge about their own users through further custom-designed surveys.

Note

1. www.prosebica.de

References

Choudhury, S., Hobbs, B. and Lorie, M. (2002), "A framework for evaluating digital library services", *D-Lib Magazine*, Vol. 8 Nos 7/8, available at: www.dlib.org/dlib/july02/choudhury/07choudhury.html (accessed March 9, 2006).

Decker, R. and Hermelbracht, A. (2005), "Pattern detection with growing neural networks: an application to marketing and library data", in Fleuren, H., den Hertog, D. and Kort, P. (Eds), *Operations Research Proceedings 2004*, Springer, Berlin, pp. 230-7.

Decker, R. and Hermelbracht, A. (2006), "Planning and evaluation of new academic library services by means of web-based conjoint analysis", *The Journal of Academic Librarianship*, in press.

Heath, F., Kyrillidou, M., Webster, D., Choudhury, S., Hobbs, B., Lorie, M. and Flores, N. (2003), "Emerging tools for evaluating digital library services: conceptual adaptations of LibQUAL + and CAPM", *Journal of Digital Information*, Vol. 4 No. 2, available at: http://jodi.tamu.edu/Articles/v04/i02/Heath/ (accessed March 9, 2006).

Hermelbracht, A. and Koeper, B. (2006), "Der Einsatz der Conjoint-Analyse in wissenschaftlichen Bibliotheken – Imagegewinn durch Dienstleistungsentwicklung 'nahe am Kunden'", in Lülfing, D. and Siebert, I. (Eds), *Deutscher Bibliothekartag in Düsseldorf 2005*, Klostermann, Frankfurt am Main, pp. 78-92.

Hermelbracht, A. and Senst, E. (2005), "Ideen für innovative Serviceangebote von Universitätsbibliotheken – Ergebnisse der Ideengenerierungsphase im Rahmen des ProSeBiCA-Projektes an der Universität Bielefeld", *Bibliotheksdienst*, Vol. 39 No. 1, pp. 50-75, available at: www.zlb.de/aktivitaeten/bd_neu/heftinhalte2005/betriebsorg0105.pdf (accessed March 9, 2006).

Hermelbracht, A. and Senst, E. (2006), "The ProSeBiCA-project: application of the conjoint analysis as a marketing research tool for the development and control of future academic library services", in Gupta, D.K., Koontz, C.A., Massisimo, A. and Savard, R. (Eds), *Marketing Library and Information Services: International Perspectives*, K.G. Saur, Munich, pp. 300-12.

Corresponding author
Antonia Hermelbracht can be contacted at: ahermelbracht@wiwi.uni-bielefeld.de

The KB e-Depot digital archiving policy

Erik Oltmans and Hilde van Wijngaarden
Koninklijke Bibliotheek, The Hague, The Netherlands

Abstract

Purpose – Electronic journals dominate the field of academic literature, and it is of great importance to the international scientific community that this electronic intellectual output remains accessible in perpetuity. The purpose of this paper is to discuss the policy and ambitions of the National Library of The Netherlands (Koninklijke Bibliotheek, KB) regarding digital archiving of electronic publications.

Design/methodology/approach – This article discusses three possible threats against permanent access, and proposes a coordinated and systematic approach to address these risks: the safe places network. This paper also includes a comprehensive overview of the e-Depot system and the KB approaches to digital preservation.

Findings – The KB e-Depot has been operational for more than three years, and fulfils the most important requirements.

Research limitations/implications – The KB focuses on both migration and emulation as preservation strategies

Originality/value – This paper fulfils an identified need for collaboration.

Keywords Digital storage, Archiving, National libraries, The Netherlands

Paper type Case study

KB policy

Virtually every country has a national (legal) deposit of printed publications, and in most cases these collections are housed in the national libraries. Gradually more national deposit libraries will also build electronic deposits for long-term preservation and permanent access. It is uncertain, however, whether the traditional model, based on national deposits and geographical frontiers, will be able to guarantee the long-term safety of the international academic output in a digital form. Academic literature is produced by multinational publishers, and has often no longer a country of origin that can be easily identified and thus no obvious guardian. Hence, in the traditional model there is a huge risk of academic records being lost forever. A systematic and more concentrated approach is needed to address this unacceptable risk.

Another threat is disrupted journal access for a certain period following a publisher failure, or publishers that stop making journals available for commercial reasons. As the prevailing model in digital publishing is licensing rather than archiving, libraries should know where they can go in case of loss, so as to guarantee continuous access for end users.

Finally, the last risk that we want to address here, is technological obsolescence. Digital material is often unstable and has a brief lifespan, because of the limited longevity of information carriers and of the software and hardware that make the stored information accessible to users. Although currently we can still render most file formats, we need to be prepared for objects appearing to be damaged or impossible to render. If we fail to pay attention and fail to continue our research efforts, this situation will be inevitable. The three major types of risks are depicted in Figure 1.

Research Libraries, End Users

Disrupted Access

Obsolescence

National Libraries

Publications without natural fatherland
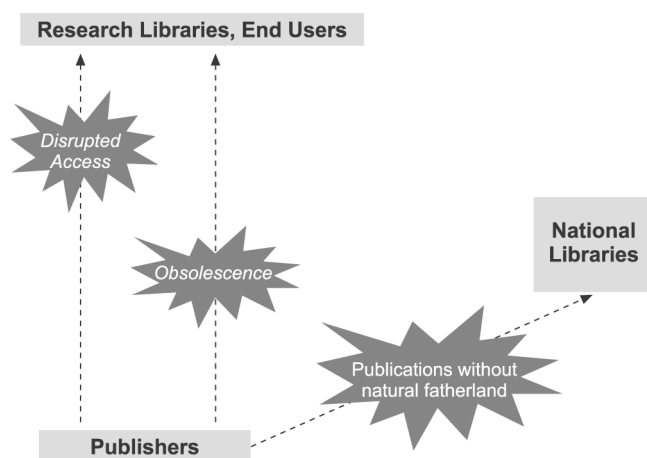
Publishers

The model that the KB proposes in order to control these risks, is called the safe places network. It is based on systematic cooperation with publishers, who deposit their materials at a limited number of safe places. We argued that many journals do not have a clear place of publication, but they do need a place to be archived safely. Publishers are not likely to deposit their material at an infinite number of digital archives. They probably want to sign archiving contracts with a limited number of institutions around the world to deposit their materials – partly to spread their risks and partly for geopolitical reasons.

These institutions, serving as permanent archives, require permanent commitment. A permanent archive should provide a reasonable guarantee of continuity. Furthermore, permanent archiving calls for substantial investment, not just financially, but also in the form of building up the necessary specific skills and expertise. Moreover, the preservation function will require an unremitting research and development commitment. From these requirements it follows that permanent archiving should be taken care of by a limited number of institutions, dedicated to this task. Permanent archiving should be prominent in their mission. Not every library should try to establish its own permanent archiving system. In the case of international scholarly journals a handful of permanent archives, wisely spread around the globe, will suffice. The economies of scale that can be achieved provide a key incentive for developing this safe place model. The initial investments that will be required, in terms of financial resources and staffing, are very high. But once these investments have been made, expanding such an operation into an international service will clearly reduce the cost per unit of stored information (see also van Drimmelen, 2004).

Many of the arguments above come from the perspective of a national library practicing electronic deposit. The recently issued statements formulated by the Andrew W. Mellon Foundation, endorsed by the Association of Research Libraries (ARL), stress other perspectives as well (ARL, 2005). They say that libraries and associated academic institutions must recognize that preservation of electronic journals is a kind of insurance against permanent loss, and that research and academic libraries may collaborate in the form of an insurance collective. Preservation is a way of managing the risk against the permanent loss of

electronic journals, and against having journal access disrupted for a protracted period following a publisher failure.

In order to address these risk factors and to provide insurance against loss, qualified preservation archives should provide a minimal set of well-defined services, storing electronic journal files in trusted archives outside the control of the publisher. Archives must receive files that constitute a journal publication in a standard form, either from a participating library, or directly from the publisher, and must store the files in non-proprietary formats. Moreover, archives should use a standard means of verifying the integrity of ingoing and outgoing files, and provide continuing integrity checks for files stored. They must also limit the processing of files, in order to keep costs down, but provide sufficient processing so that the archives could locate and adequately render files for participating libraries in the event of loss. And finally, archives must restrict the access of the participating libraries to archived files that are under copyright, in order to protect the publisher's business interests, except when the publisher goes out of business or is otherwise unable to provide consistent access. These trigger events would be the main exception allowing widened access to third parties like other libraries or end users (see Figure 2).

Whether the concept is called "safe places network" or "insurance collective", the implication is clear: dedicated institutions are needed to take on the responsibility. The KB aims to play a prominent role within the international safe places network. It has implemented all the requirements mentioned above: with the operational digital archive e-Depot, the KB possesses a sound technical and organizational infrastructure and specialist skills and expertise, and has committed itself to an ongoing research and development effort. These assets provide a firm foundation on which to expand the e-Depot's international role, generating substantial economies of scale, since it enables the investments necessary for the national e-Depot to be used even more efficiently.
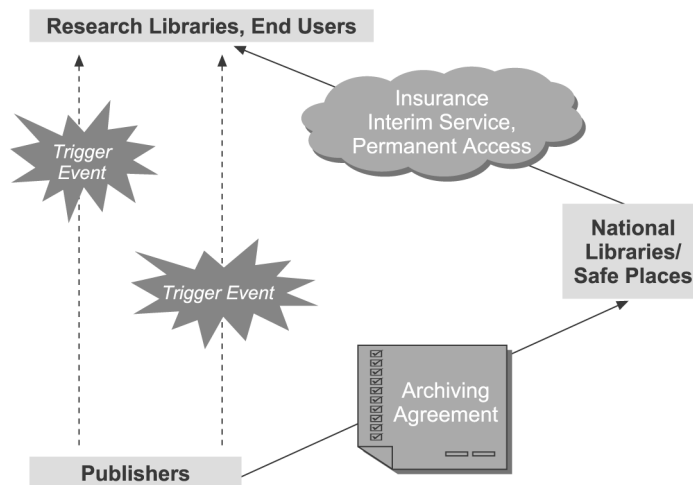
Figure 2.
Addressing the risks
against permanent loss

Governance, funding and organizational structure

The KB was founded in 1798 and since 1993 has been an autonomous administrative body financed by the Ministry of Education, Culture, and Science. The KB receives an annual grant from this Ministry, amounting to €40 million in 2006. The KB also has some internally-generated income (library passes, document supply and interest), which amounts to less than 10 percent of the annual budget. The KB may apply for additional funds to support special projects or investments in the infrastructure.

As for the e-Depot, the KB has re-allocated funding within its own budget for several years. In addition, since 2003 the KB receives an earmarked grant of €1.1 million per year from the ministry for system maintenance and for the staff handling the operations of the e-Depot. The system maintenance is outsourced to IBM. The associated research and development budget was an additional €200,000 for staff. In 2005 this annual grant went up to €0.9 million, exclusively dedicated to research into digital preservation. These funds are expected to increase further in 2006 and 2007, subject to approval by the cabinet.

The e-Depot system falls under the Acquisitions and Processing Division, whereas the Research and Development Division includes the department for Digital Preservation research. The IT division is responsible for technical maintenance, together with IBM. The total number of staff handling the system, ingesting the publications, research projects, and management, is equivalent to more than 15 full time posts.

Agreements with publishers

In 1993 the KB decided to build a deposit collection of electronic publications, which was a logical extension of the deposit collection of printed publications already in place. General policy lines were formulated, and in 1995 the KB started experimenting on a small scale with facilities for automatic handling of e-publications.

With this extension of tasks, the KB was confronted with the dilemma of electronic media: its short life expectancy. Digital material has a brief lifespan, because of the limited longevity of information carriers and of the software and hardware that make the stored information accessible to users. Therefore, since 1994 research and development on long-term digital preservation has been a topic of growing importance for the KB.

In 1996 the KB and the Dutch Publishers Association agreed on an arrangement for the voluntary deposit of offline electronic publications. At the same time, discussions were initiated with Elsevier Science aiming at acquiring the content of Elsevier e-journals with Dutch imprint, and the first experimental bilateral archiving agreement was signed. Soon afterwards a similar archiving experiment was agreed on with Kluwer Academic (see also Steenbakkers, 1999). The Dutch Publishers Association agreed on a new arrangement in 1999, which covered offline as well as online electronic publications with Dutch imprint (updated again in 2005).

A landmark electronic archiving agreement was drawn up with Elsevier Science in 2002: the experimental agreement of 1996 was expanded to cover the entire set of Elsevier journals. In total, the agreement defined the responsibility for preserving nearly all Elsevier journals, also covering journals digitized as part of Elsevier's retrospective digitization project (estimated at a total number of approximately seven million articles). This arrangement turned the KB into the first official digital archive in

the world for journals published by an international scientific publisher. In 2003 an official archiving agreement with Kluwer Academic followed. The early and successful implementation of the e-Depot and the commitment of Kluwer and Elsevier, based on trust and commercial interest, put the KB in a natural position to assume an international role. After the agreements with Elsevier and Kluwer, the KB concluded similar agreements with:

- BioMed Central (2003);
- Blackwell Publishing (2004);
- Oxford University Press (2004);
- Taylor & Francis (2004);
- Sage Publications (2005);
- Springer (2005); and
- Brill Academic Publishers (2005).

The third publisher the KB entered into an agreement with was BioMed Central. This contract signified an important step in two ways. Firstly, it underlined the international role of the national deposit system. BioMed has no Dutch origin. Secondly, BioMed was established as an open access publisher right from the start. This also was new to the KB. Thus, the BioMed agreement represented a major strategic step. As the list of publishers makes clear, the KB does not discriminate between the places of origin, the publisher's business model, marketing strategy or any other features.

Designated community
There is a minimum set of conditions to be fulfilled before the KB enters into an archiving agreement. Publishers must deposit their publications free of charge. However, the KB has to accept restrictions on access, avoiding interference with the publisher's commercial interests. But there is a minimum level of provision: the KB provides permanent access to the journals on site to all authorized library users, including availability for interlibrary document supply within The Netherlands, and including remote access if allowed by the publishers. For example, the archiving agreement with BioMed Central secures free remote access to over 100 open access journals covering all areas of biology and medicine. In addition, should there be a catastrophic event, such that the publisher is inoperable for a long period of time, the KB would be part of the interim service system. The official archive thus serves as a guarantee to all licensees worldwide, by safeguarding the access that licensees have paid for. Finally, should the publisher or a successor cease to make these journals available, the KB could open access to all on a walk-in or remote basis. In this way, the KB secures permanent access to both libraries and end users, without threatening a publisher's business interests.

Content characteristics
The e-Depot's content is predominantly driven by the archiving agreements. At present the e-Depot is receiving two types of electronic publications: offline media (CD-ROMs that are fully installed before they are loaded into the e-Depot, including

operating systems and additionally required software) and online media such as the electronic articles deposited by publishers. In March 2006, the e-Depot contained over 5.8 million digital objects, corresponding to a little more than 6 terabytes of storage space. The total number of e-journal titles is over 3,500. Full implementation of all current archiving agreements will result in an electronic archive containing more than 9 million digital publications. The annual increase in the number of articles from these publishers will be around 400,000.

The aim of the KB for the coming years is twofold. The KB will actively try to conclude archiving agreements with more of the major international scientific publishers. The twenty largest publishing companies cover almost 90 percent of the total world production of electronic science-technology-medicine (STM) literature and the KB would like to reach that level of coverage in the e-Depot. The KB will also try to obtain the most cited scientific journals for its e-Depot, irrespective of the publisher. Alongside this active strategy, the KB will accept electronic literature from any other publisher who wishes to deposit material with the e-Depot, provided that the publisher is able to deliver the material in the preferred format and with the necessary metadata, and provided the publisher complies with the minimum set of access conditions as stated earlier.

Apart from archiving scientific digital publications, setting up a successful archiving workflow and infrastructure has also opened up opportunities for the long-term storage of other kinds of digital material. Projects have started to develop functionality and models for the storage of digitized material and websites. The KB is also working together with the Dutch university libraries to store their scientific output for the long-term in the project digital academic repositories (DARE)[1].

Technical architecture and workflow

The first experimental deposit system was based on AT&T Right Pages. When Right Pages was withdrawn from the market in 1996, IBM Digital Library was selected to replace the AT&T software. It was recognized that IBM Digital Library was only a temporary solution because it did not have the functionality needed for a full-scale deposit system. In 2000, after a European tender procedure, IBM was selected to develop a new system together with KB staff. In this project the expertise of the KB and the technical knowledge and research forces of IBM were combined, resulting in digital information and archiving system (DIAS). In late 2002 DIAS was delivered and embedded, resulting in the current e-Depot system. It is now fully operational and embedded in the KB organization, as a department within the Acquisitions and Processing Division. As well as at the KB, the DIAS system is also in use at Die Deutsche Bibliothek (DDB), the German national library. The current users of the DIAS system meet twice a year to exchange experiences and to work together on improving the functionality.

The infrastructure of the e-Depot consists of both components that were specifically developed for processing, archiving, and maintaining e-publications, and typical digital library functions. According to the Networked European Deposit Library (NEDLIB) guidelines, the deposit system should be a separate, dedicated entity within the library's digital infrastructure. For the traditional library processes, such as cataloguing, search and retrieval, and user registration and authentication, the KB uses the provisions already in place, thus avoiding duplicating these functions within the

deposit system. This approach allows both the e-Depot system and the traditional library systems to evolve at their own pace.

The installable CD/DVD-based publications are firstly completely installed on a Reference Workstation, including all additionally required software such as image viewers and media players. A snapshot of the fully installed publication – together with the operating system on which it is installed – is then generated into a disk image. For these electronic publications, it is the disk image that is ingested into the e-Depot, and patron use requires retrieving the disk image and completely installing it onto a workstation (Oltmans and van Wijngaarden, 2004).

Most electronic publications and their associated files are obtained via digital tape or are acquired via file transfer protocol (FTP). The files are validated first, and then batched for further processing, while corrupt content is recognized automatically and is dealt with according to error handling procedures. The processing ingests both the content files and the metadata. It converts the publisher's bibliographic data into the KB's standard format and adds a national bibliographic number (NBN) which is later used as the unique identifier of the stored item. There are functions for search, retrieval, and delivery: the local overall catalogue database is freely available, whereas the content itself is only available after a procedure for identification, authentication, and authorization (IAA).

The functional design of DIAS is based on the open archival information system reference model (cf. Consultative Committee for Space Data Systems, 2002). The system is designed to be durable, and provides for scalability and flexibility. In 2003 an international task force on digital repository certification was initiated by the Research Libraries Group (RLG) and the National Archives and Records Administration (NARA)[2], which has developed an audit tool which is now being tested. Its purpose is to produce certification requirements for establishing and selecting reliable digital information repositories. Three digital archives have been selected as pilots for the test-audit, of which the KB e-Depot is one. The test-audits are taking place in February-April 2006 and will involve the investigation of the e-Depot's organizational and technical infrastructure and processes. The purpose of auditing a digital archive is to determine the degree of certainty the archive provides for the long-term availability and the functionality of the digital resources that are stored. The audit should ultimately result in a certified system.

The KB approaches to digital preservation

Providing permanent access to electronic material is a complex problem. As has been said, digital material is often unstable and has a brief lifespan, because of the limited longevity of information carriers and the software and hardware that make the stored information accessible to users. Safeguarding the integrity and authenticity of the material is therefore a key challenge when dealing with long-term preservation. Regardless of the chosen strategy, permanent access calls for continuous attention and action. The rapid pace of technological change means that the techniques and procedures for long-term storage and accessibility requirements need to be adjusted and improved constantly. A permanent R&D effort is therefore indispensable.

There are two main approaches to digital preservation. The first one is migration and focuses on the digital object itself. It aims at changing the object in such a way that software and hardware developments will not affect its availability. By changing or updating the format of an object, it is made available on new software and hardware.

The digital object will be adjusted to changes in the environment, which makes it possible to render objects using current systems. The second approach is emulation, which does not focus on the digital object itself, but on the environment in which the object is rendered. It aims at (re)creating an environment in which the digital item can be rendered in the same form as upon delivery to the archive.

There are arguments for preserving the original look and feel, as well as for converting documents to new standards. Both models are therefore being studied and considered for implementation at the KB, taking cost issues into account (cf. Oltmans and Kol, 2005).

The main reason for preserving the authentic form is that the KB digital archive serves as a safe place for original materials from publishers. The KB promises to do its utmost to safeguard the integrity of the articles that are deposited in the e-Depot. If at all possible, KB wants to save an article "as is". In digital preservation the contradiction is that changing less, implies doing more. In the long term, emulation tools will be needed to render these publications in the same way as they were published originally, and this kind of emulation tool does not exist yet. The development of preservation-based emulation is also important for those end users who want to access publications and experience the original look and feel. In the shorter term, migration can be performed without changing the content too much, especially when considering plain text articles. But if migrated articles have to be migrated again, stacked errors may occur, which damage the integrity over time. On top of this, new publishing formats offer the opportunity of adding moving images, interactive models or spreadsheets, challenging current migration techniques. In this situation, emulation becomes a necessity rather than a choice. This is why KB has started a project to develop a modular emulator for digital preservation together with the Nationaal Archief of The Netherlands (van der Hoeven and Van Wijngaarden, 2005). In 2007 the emulator will be delivered for implementation in the e-Depot infrastructure.

However, emulation is certainly not our only strategy. Migration can be a good alternative in the short term, as is said above. It can also be developed further with the aim of offering future access to publications according to the standards and functionalities of that time. Migration will be needed to enable future browsing, copying and reusing data. Therefore, in 2006 KB began a project to do research on the quality of existing migration tools and on the way these tools can be integrated in the KB workflow. Gaps in the availability of viable migration tools will be identified, resulting in new plans to fill these gaps.

In order to execute preservation plans we also need structured information about the technical properties of the stored file formats. Together with IBM, an application for the storage of technical metadata and rendering information has been developed, which is called the "preservation manager" (Oltmans *et al.*, 2004). In 2006, KB will also implement the JSTOR/Harvard Object Validation Environment (JHOVE), an application developed by Harvard University Library and JSTOR to extract technical information from delivered publications[3].

The development of preservation planning capability requires a permanent R&D effort, focused on the full range of available preservation techniques. The KB has developed its own R&D-program for the coming years and acknowledges the need for international collaboration which will result in widely accepted technologies, preferably in distributed environments. An important step towards joining forces in

this area is the European project Planets (Preservation and Long Term Access through Networked Services). Planets was proposed to the European Commission in the 6th European Framework Program for Research and Technological Development and is currently under negotiation to start in May 2006. Planets is coordinated by the British Library and brings together a diverse but expert group of partners. Stakeholders in finding solutions for preservation planning and permanent access are the national archives of the UK, The Netherlands and Switzerland and the national libraries of Austria, Denmark, the UK and The Netherlands. Research institutes are working on the problem from a more scientific point of view and will work together with the stakeholders on building solutions. Partners are the University of Cologne, University of Glasgow, University of Freiberg and the Technical University of Vienna, which all have experience and expertise in the field. Technology vendors have joined our group to build tools and the technical infrastructure that will allow us to set up work together in a networked environment. These technical partners include the Austrian Research Center, IBM, Microsoft and Tessella. The participation of these commercial partners facilitates the take-up and dissemination of research results. The result of the project will be a distributed preservation framework for the development and application of tools for preservation planning, preservation actions (tools) and content characterization. It will also include a decision support system, which will help institutions to decide which preservation strategy suits their situation best.

Summary

The KB's policy and ambitions regarding permanent archiving of electronic publications can be summarized as follows:

- There is a growing volume of electronic publications without a natural country of origin which are crucial for academic research.
- These publications must be preserved for the long term, by organizations who take on the responsibility, and who are dedicated and equipped for this task (safe places).
- The KB has the ambition to be one of these safe places, and has had an electronic deposit system in place for nearly three years; its policy is acknowledged by the government.
- The KB looks forward to concluding archiving agreements with more international publishers.
- Two prominent methods for permanent preservation are being studied and implemented, in close collaboration with international partners.
- The KB is constantly seeking opportunities for collaboration, and would like the e-Depot to be audited by an independent organization, preferably according to ISO-certification procedures.

Notes

1. www.darenet.nl
2. www.rlg.org/en/page.php?Page_ID = 367
3. http://hul.harvard.edu/jhove/

References

ARL (2005), "ARL endorses call for action to preserve e-journals", October 31, available at: www.arl.org/arl/pr/presvejrnloct05.html (accessed April 19, 2006).

Consultative Committee for Space Data Systems (2002), "Reference model for an open archival information system (OAIS)", *Blue Book*, January, available at: http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed April 24, 2006).

Oltmans, E. and Kol, N. (2005), "A comparison between migration and emulation in terms of costs", *RLG DigiNews*, Vol. 9 No. 2, available at: www.rlg.org/en/page.php?Page_ID=20571#article0 (accessed April 19, 2006).

Oltmans, E. and van Wijngaarden, H.N. (2004), "Digital preservation in practice: the e-Depot at the Koninklijke Bibliotheek", *VINE, The Journal of Information and Knowledge Management Systems*, Vol. 34 No. 1, pp. 21-6.

Oltmans, E., van Diessen, R.J. and van Wijngaarden, H.N. (2004), "Preservation functionality in a digital archive", in *Proceedings of the 4th ACM/IEEE-CS joint Conference on Digital Libraries, Tucson, Arizona, June 11*, ACM Press, New York, NY, pp. 279-86.

Steenbakkers, J.F. (1999), "Developing The Netherlands depository of electronic publications", *Alexandria – The Journal of National and International Library and Information Issues*, Vol. 11 No. 2, p. 93.

van der Hoeven, J.R. and Van Wijngaarden, H.N. (2005), "Modular emulation as a long-term preservation strategy for digital objects", in *Proceedings of the 5th International Web Archiving Workshop (IWAW05) Held in Conjunction with the 8th European Conference on Research and Advanced Technologies for Digital Libraries (ECDL 2005)*, September 22-23, Vienna, available at: www.iwaw.net/05/papers/iwaw05-hoeven.pdf (accessed April 19, 2006).

van Drimmelen, W. (2004), "Universal access through time: archiving strategies for digital publications", *Libri, International Journal of Libraries and Information Services*, Vol. 54 No. 2, pp. 98-103.

Corresponding author
Erik Oltmans can be contacted at: erik.oltmans@kb.nl

Bielefeld Academic Search Engine (BASE)

An end-user oriented institutional repository search service

Dirk Pieper and Friedrich Summann

Bielefeld University Library, Bielefeld, Germany

Abstract

Purpose – The purpose of this paper is to describe the activities of Bielefeld University Library in establishing OAI based repository servers and in using OAI resources for end-user-oriented search services like Bielefeld Academic Search Engine (BASE).

Design/methodology/approach – Discusses OAI based repository servers.

Findings – BASE is able to integrate external functions of Google Scholar. The search engine technology can replace or amend the search functions of a given repository software. BASE can also be embedded in external repository environments.

Originality/value – The paper provides an overview of the functionalities of BASE and gives insight into the challenges that have to be faced when harvesting and integrating resources from multiple OAI servers.

Keywords Search engines, University libraries

Paper type Technical paper

In a position paper of the Scholarly Publishing and Academic Resources Coalition (SPARC) published in 2002, Raym Crow (2002) defined an institutional repository as a "digital collection capturing and preserving the intellectual output of a single or multi-university community" (p. 4). Repository servers can help institutions to increase their visibility and, in addition, they are changing the system of scholarly communication.

For several years libraries have been facing price increases for scientific journals, which has led to the fact that the proportion of published journal articles libraries can provide access to has decreased. Aside from the necessity of developing new subscription models for journal articles, libraries now can do a lot to increase the availability of journal articles by providing access to open access journals and documents and building repositories for their home university.

In addition to several self-designed repository based services, Bielefeld University Library has developed, with support from the Norwegian company Fast Search & Transfer[1], an end-user-oriented search service for multiple scholarly full text archives, digital repositories and preprint servers on the worldwide web called Bielefeld Academic Search Engine (BASE)[2] (see Lossau, 2004; Lossau and Summann, 2004). At the time of the 8th International Bielefeld Conference in February 2006 BASE contained about 2.7 million documents in 189 collections. An up-to-date overview, including the content providers, is available online in a comprehensive list[3]. Characteristics of BASE include:

- intellectual selection of resources;
- indexes contain only quality-assured academic online resources from all academic disciplines;
- transparency about the data resources included in BASE;
- searches metadata and full text (depending on the data source);
- discloses internet resources of the "deep web" (such as 500,000 digitized pages of historical journals and review organs of the German enlightenment);
- displays search results as bibliographic data and full text hits;
- various options to sort result sets; and
- search refinement for authors, keywords, document type, or language.

The newest feature of BASE, which was first presented to the public at the 8th International Bielefeld Conference, is the ability to check BASE results in Google Scholar by a title search, so that users can directly see if, and how many times, an article is cited in Google Scholar. Figure 1 shows the result of a simple search for "hawking radiation". After clicking on the link "Check this title in Google Scholar" for the second hit, a window with the result in Google Scholar pops up.

The integration of a citation counting functionality in institutional repository servers is something for which there is a high demand from academics. So the basic idea behind this feature is that while our search engine software, like most repository software, does not provide this yet, there is no reason why an external system like Google Scholar should not be used for this functionality. BASE is also flexible enough to combine data collections in special views, e.g. for all institutional repository collections. It is also possible to replace or to amend the search functions of a given repository software, which we will demonstrate in the near future when establishing an e-scholarship repository server for Bielefeld University.

Regarding the technical background of repository based services, Bielefeld University Library is working on both sides – establishing OAI services and using those of other institutions. It has been running the publications server Bielefeld Server for Online Publications (BieSOn) since 2004 and Bielefeld Text Archive Server (BieTAS), a platform for the comfortable dissemination of distributed contents under different systems, since 2005. Both services are registered OAI servers. The e-scholarship repository server mentioned above will soon follow. On the other hand, collecting metadata via OAI harvesting plays an important role in different search environments. We feed more than 550,000 documents into the local library catalogue as online-accessible material. This includes documents such as theses and dissertations, digitized books and journals. A total of 2.4 million journal articles, harvested for example from Citeseer, PubMed, ArXiv, the Directory of Open Access Journals (DOAJ), and Biomed Central, are added to the local article database as electronic references. The most relevant dataflow is feeding all type of material (images, maps, videos, multimedia components and web pages) into BASE. In addition to the crawling of web pages, the data processing of OAI metadata has become the main focus of the BASE data workflow. To process this data we have established a pre-processing stage to transform OAI metadata Dublin Core XML files into an internal XML format. As the next step these files are transformed by a series of different internal and external processing stages into a file which can be indexed directly afterwards. This includes, in
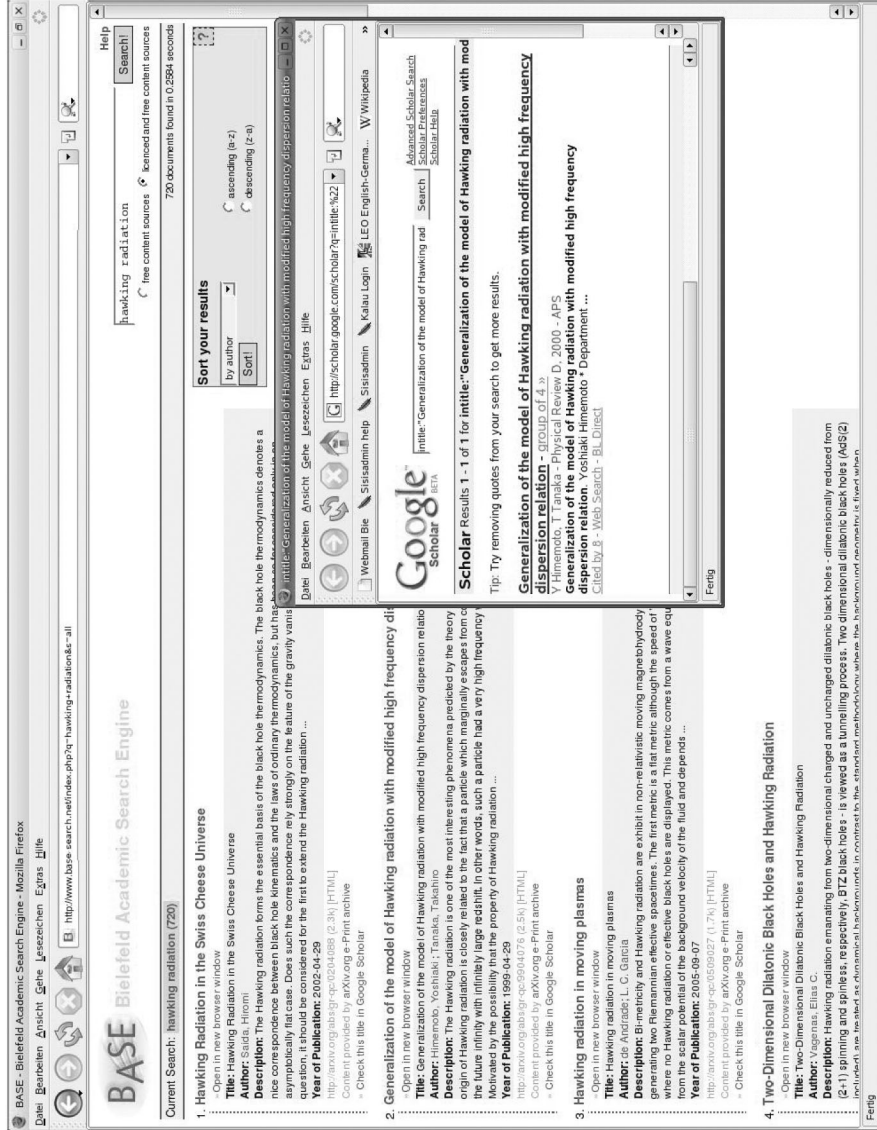
Figure 1.
Checking citations of
BASE search results in
Google Scholar

particular, format transforming (e.g. PDF, PS, ZIP or Office files), language detection, normalizing and lemmatizing. After indexing, all this information is ready for retrieval – in our case for accessing it via the BASE search interface based on PHP scripts.

Focusing on the harvesting process, the first challenge is how to find relevant academic OAI servers. We monitor the well-known registries of openarchives.org, Eprints, the experimental registry of the University of Illinois, DSpace, and, since January 2006, the Directory of Open Access Repositories (DOAR). These resources provide different numbers of listed servers and a different quality of stability and status of their data. The map in Figure 2 shows the geographical distribution of repositories covered in BASE with the main focus on academic repositories in Europe. The map illustrates the strong position of Sweden, Germany, the UK and The Netherlands. Additionally BASE integrates a large number of repositories from the US, accompanied by Australian and Canadian repositories.

The harvesting procedure itself proved to be complicated in detail and posed a number of problems. To handle these problems, and to make the process more efficient, we adopted and developed a small collection of software tools. Firstly, as our core system we are using the Perl-based open source harvester delivered by the US company FS Consulting. While harvesting we faced some minor error situations that we were able to solve by adapting the source code. Relatively often the delivered OAI data contains XML errors. This is a serious problem to deal with, because XML parsing is then impossible for the whole file. Therefore, we wrote an XML validator and repairer script which removes the invalid records and saves the correct ones. A so-called Harvest Watcher monitors the harvesting processes and reports the results (count of records, time stamp). A cronjob script, the OAI Resource Updater, automatically requests repository servers in defined intervals. Finally, the Registry Watcher takes the valuable HTML or XML files which are delivered from some of the registries, compares them with the BASE harvesting configuration file and lists the

Figure 2.
Institutional repositories
covered in BASE

servers already covered, and, much more interestingly, the resources unprocessed up to now. All these additional tools have been written in Perl.

While the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) defines the OAI harvesting process very clearly, the daily routine brings up a broad set of serious problems and challenges. Among the long list of problems are non-responding servers, document links that do not work, invalid XML files and OAI data which only contains references either without any full text behind or where full text access is restricted for specific access. This situation requires a lot of observation and a lot of detailed configuration work in response. Some short examples will deliver a deeper insight into the problems. Some installations deliver the URL for access in the "source" field. Sometimes authors' names are inserted as a list with different separators in one Dublin Core (DC) "creator" field. On the other hand, one can also find author names split into two different fields. The content of the DC field "subject" contains classification codes, classification terms or true subject headings without any qualifying. In some cases one can find author and title information in this field as well. A bitter experience is that fields with rather standardized content, such as the DC fields "date" and "language", vary in a very broad way. In particular, the "language" field should determine the language of a resource correctly, because this is the basis for several linguistic processing steps. Another significant quality problem is the fact that correct citation information for journal articles is missing among many OAI servers. To handle all these problems it has been necessary to put a lot of effort into registration and configuration. This has taken more time than expected.

As a conclusion of our experiences we can establish a list of personal rules derived from the harvesting activities. Firstly, standard repository software is very useful, not only for the work of system administrators, but for the OAI harvesting procedure as well. The delivered results of those systems are strongly standardized, which makes the integration process much easier. Besides that, small collections generally only bring up small problems, probably because the content is more basic and more homogeneous. A serious problem is combining metadata and corresponding full text via OAI because the linking method and the presentation of the documents vary in practice. This is the reason why we only succeeded in realizing this approach for a few installations. Another point shown by experience is that libraries as data providers produce a higher level of data quality, probably because they have much broader experience with bibliographic metadata. An important point for improving the quality is participating in the OAI community. Writing e-mails to the repository administrators helps, but sadly only sometimes. In 60 percent of our e-mail contacts we got a response, and in half of those cases the problems were repaired within a week. Sometimes there was no response at all but some weeks later the problem was solved, perhaps an internal reaction to the e-mail contact. As a last point we have to mention that OAI data aggregation, and, in particular, using aggregator services, may produce problems. In some cases we faced duplicates, updating delays and loss of individual information on the way from the original repository to the aggregator service.

As a last topic, the aspect of integrating BASE in other services has necessitated an ambitious approach to technology, especially the idea of embedding the system in external repository-based environments. The easiest way of integration is to include a search form for BASE. This feature works already and there is HTML code available that can be incorporated into any user-defined web page. In cooperation with a German

library software company we have developed an HTTP-based interface to integrate BASE retrieval in a more flexible way. This technique will be improved with a more comfortable interface. Besides this, we are working on a web services-based technology that accepts and responds with XML files, including search queries and result pages. In relation to the German project Vascoda we have discussed a concept for a federated search of different search engines with a high level of result-merging based on IT standards. Hopefully this approach will support the development of another type and quality of search environments in the future.

Notes

1. www.fastsearch.com
2. www.base-search.net
3. http://base.ub.uni-bielefeld.de/about_sources_english.html

References

Crow, R. (2002), "The case for institutional repositories: a SPARC position paper", available at: www.arl.org/sparc/IR/IR_Final_Release_102.pdf (accessed May 9, 2006).

Lossau, N. (2004), "Search engine technology and digital libraries, libraries need to discover the academic internet", *D-Lib Magazine*, Vol. 10 No. 6, available at: http://dx.doi.org/10.1045/june2004-lossau (accessed May 9, 2006).

Lossau, N. and Summann, F. (2004), "Search engine technology and digital libraries: moving from theory to practice", *D-Lib Magazine*, Vol. 10 No. 9, available at: http://dx.doi.org/10.1045/september2004-lossau (accessed May 9, 2006).

Corresponding author
Dirk Pieper can be contacted at: dirk.pieper@uni-bielefeld.de

The Nereus international subject-based repository
Meeting the needs of both libraries and economists

Vanessa Proudman
Tilburg University Library, Tilburg, The Netherlands

Abstract

Purpose – The purpose of this article is to provide an overview of the essential partnerships needed to better guarantee strong added value institutional repository (IR) service provision for faculty user satisfaction, resulting in the essential content acquisition for IRs.

Design/methodology/approach – Three crucial partnerships are described as a basis for the development of a meaningful institutional repository service. This paper concentrates on two partnerships: the first describes the importance of partnership between institutions; in this case, an international subject-specific network: Nereus is an innovative consortium of 15 prestigious European universities and institutes in the area of economics. The second partnership goes into that of the library with its IR content providers and service end users by taking the Economists Online (EO) project model as a case in point.

Findings – University libraries and their changing active roles are building valuable partnerships with their clients and information providers; the Nereus consortium, and its EO project is a case in point where benefits are felt by both libraries and researchers. EO is a Nereus cornerstone and aims to increase the usability, accessibility and visibility of economics research by digitizing, organizing, archiving and disseminating the complete academic output of some of Europe's leading economists, with full text access as key. It is building an integrated online showcase of Europe based on IRs. EO's prime goal is to focus on services of direct value to the author, e.g. providing new full text content online, digitizing older material, creating automated publication lists, metadata quality-control, more focused dissemination of content – all from one repository source and complimentary to the services publishers provide. Successful partnerships can be formed by designing and offering a strong product and service, combined with a good advocacy program containing arguments which support the researcher in his/her work process, and addressing real problems.

Originality/value – Nereus and Economists Online have the potential to be a model for adaptation in any disciplinary sector for disseminating and giving heightened access to scientific output online. Some of the critical success factors of these initiatives have been identified and expanded upon in the paper.

Keywords Academic libraries, Value added, Incentive schemes, Europe

Paper type Case study

Introduction

Libraries are repositioning themselves in a changing world of information retrieval and publishing. National and international funding programs are on the increase, promoting R&D, institutional repositories (IRs) and above all the utilization of networks for cost-efficiency and knowledge exchange. The deployment of institutional repositories is a key strategic aim for many countries where libraries now have further potential to help advance scholarship, by supporting the author in his/her research and

by delivering added value, quality services through IRs. The focus needs to be far more on the individual information provider's, i.e. the author's, work processes and needs, in order to ensure success and sustainability. Once this has been achieved, the library's aim to increase the acquisition of high-quality institutional output in an IR is also guaranteed. It is above all the existence and nurturing of trusted partnerships with key stakeholders that is essential: be this with the author, management or the institutional network. Such partnerships form the backbone of any strong project idea, allowing it to fully come to fruition.

What kinds of models or projects should therefore be sought in order to achieve the quality and critical mass of content so essential to both the institution and the individual researcher? The Nereus network and program and its project Economists Online is one such a model; I will share our experiences in this paper.

Three essential partnerships
Three vital partnerships need to be in place to ensure the development, acceptance and sustainability of an IR project of any significance to the international researcher:

- partnership between the university, faculty and library management;
- partnership with the international research community at an institutional level (the Nereus network); and
- partnership with the author and reader.

I will develop on the last two partnerships in particular in this paper.

Partnership 1: management
As a first step, it is important to maximize partnership with university management (university boards, rectors, library boards, etc.) when approaching repository work; the new strategic choices and opportunities open to universities in the area of scholarly communication need to be crystal clear. Once convinced, the university management can then serve to underpin the initiative's importance for the university as a whole, expressed in either material or non-material terms, thereby bringing with it crucial support from faculties. In the ideal case, the IR is made a strategic aim for the university, as at Tilburg University, The Netherlands. Thus the IR can be embedded in university activities and the foundations for a more sustainable future can be laid.

At other management levels, Faculty Deans and Faculty Heads of Research and/or Administration are crucial for the support from the research community, for know-how on current trends and work process problems, for integration with research management systems where possible, and for enhanced access to the individual contributors to the future IR project. This partnership is particularly essential for partnerships with the author and reader (partnership 3, see below).

Lastly, it is clearly evident that to achieve the IR's operational aims and get the necessary results, and to embed it efficiently within the organization, the backing of the library management is a pre-requisite. Added value for the library and its improved partnerships with stakeholders will benefit the library in the future.

Partnership 2: international subject-specific network: Nereus – institutional level
Recent developments have shown that formalized networks in the area of ICT and research, with efficiency and innovation as an aim, are on the increase. Take the recent

international "knowledge exchange" group – a network of national funders of higher education infrastructure programs such as the German Research Foundation (DFG), SURF, the Joint Information Systems Committee (JISC) and Denmark's Electronic Research Library (Deff) – as an example. A formalized partnership amongst a network of institutions can increase individual efficiencies by sharing capacities and optimize the exchange of best practices for a common goal. However, to my mind, what needs more of a place in the library community is the subject-specific network. This is frequently seen in various forms in the research community, but libraries can take this as an opportunity to meet common aims with common users at an international level. Nereus is one such an example, that aims collectively to deliver better services of international significance to the researcher, teacher and student in a way that individuals or their institutions cannot.

The Nereus consortium and program was established as a network of university libraries specialized in the area of economics in mid 2003, as a result of the successful EU library projects Decomate and Decomate II. The Nereus consortium currently (as of March 1, 2006) consists of 15 university and institutional libraries, from eight European countries including France, Germany, Spain, The Netherlands and the UK. Current members are: Carlos III University of Madrid (E), Charles University, CERGE-EI (CZ), Erasmus University Rotterdam (NL), Katholieke Universiteit Leuven (B), London School of Economics (UK), Maastricht University (NL), Sciences Po, Paris (F), Tilburg University (NL), Université Libre de Bruxelles – ULB (B), UCD Dublin (IRL), UCL (University College London) (UK), University of Oxford (UK), University of Warwick (UK), Vienna University of Economics and Business Administration (A) and ZBW – German National Library of Economics (D).

To be eligible for membership all member institutions need to have a leading research capacity in the area of economics and/or need to hold a particularly strong economics collection such as the German National Library of Economics (ZBW Kiel). All members feature in the European institutions of the ranking report based on publication output and citations commissioned by the European Economics Association[1] and produced by Tom Coupé (2003) of ECARES, which is used as a guide for membership acquisition.

Nereus is a membership organization where members pay an annual fee of €5,000 and invest in-kind with 0.25 mje (man year equivalent) staff time in order to pursue an array of Nereus activities in the Nereus rolling plan. The management structure is made up of two key sets of actors:

(1) A steering committee representing all members of mainly library directors or department heads: the strategic arm of the network.

(2) An operational group, also with complete membership representation, where information and IT specialists collaborate on developing and delivering concrete products and services for both the network and its economists.

A scientific advisory board is a sounding board for Nereus strategic planning and progress, made up of leading European economists. Nereus's current Program Director is the Librarian and Director of IT Services at the London School of Economics with the current Program Manager and Financial Director at Tilburg University.

Nereus's central motivation is on the one hand to work collaboratively to facilitate professional knowledge exchange for innovation in the library community. On the

other it is determined to showcase first-rate research in one specific international subject community through the development of added-value online information services for the economist. Nereus thereby serves three distinct user groups: the service provider (library), the information provider (the author) and the end user (economics researcher, teacher and student).

Nereus aims to serve its academics by increasing the global visibility of European excellent economics research, with a view to opening up to the USA before long. Through initiatives such as Economists Online efforts are being made to make content more visible in the information services of importance to the economist such as Research Papers in Economics (RePEc), Social Science Research Network (SSRN), Google, Google Scholar, via the Nereus repository. Similarly, Nereus is developing added value services that are of a direct benefit to the researcher. These span from unlocking new content (open access traditional publications, with a view to including datasets in the future), to creating dynamic publication lists for Nereus's leading economists with full-text links, to a resource discovery tool by way of the Nereus international index of licensed economics resources for example. For a further insight into our activities, see our rolling activity plan[2].

Nereus serves its network by focusing on cost-efficiency and innovation and by utilizing expertise from a number of libraries, from IT to marketing to content specialties. It has collectively carried out user studies on information behavior as well as on version management, and has exchanged experiences and concluded with cost-effective solutions to operational problems, e.g. an advocacy toolkit for institutional repositories, overviews on repository software choices and challenges. Service infrastructure development issues which have been addressed are: standard-setting such as metadata standards, quality issues relating to repository content, as well as technical issues surrounding portlets, interoperability, authentication (Shibboleth) and linking technologies such as Open URL.

The Nereus network serves to increase the possibilities whereby R&D work can be carried out collaboratively on projects such as Economists Online (see below)[3], and Versions[4], as well as on new EU tenders which individual institutions or small groups could not hope to achieve on their own.

In addition, Nereus pools its resources and skills for efficiency, e.g. by researching the self-archiving policies of approximately 200 publishers of international importance to the economist. Other examples of cooperation are the exchange of best-practices through product development and focused workshops, for example in the area of repositories in the context of Economists Online, as well as in other areas such as portals, technology, marketing, and authentication.

Nereus has a rolling three-year activity plan, with concrete deliverables. This forms the basis of a program that has concrete products to deliver. Activities include an international index of economics resources, the development of an index of datasets for economics, workshops on repository issues such as challenges between research information systems and repositories, as well as data librarianship.

In summary, the Nereus consortium and network of high-ranking institutions is an international partnership and infrastructure which can more efficiently and effectively serve both its own aims for building repositories, its services and content, and those of the international research community of economists, whose demands on libraries have no institutional boundaries and who want international quality content and exposure.

Partnership 3: author and reader – individual level
An institutional repository's success is clearly very much dependent on the information provider's cooperation in the process. In a library community, where user acceptance is increasingly important to justify budgeting and sustainability, where library action rather than reaction is called for, and where supporting the researcher in his/her work process can offer new opportunities to libraries, project and service design needs to have the content provider and end user even more central than previously. The Nereus pilot project Economists Online (EO) has been designed with the interests of the international economics community in mind, profiling some of its partners' individuals, and offering them new added value services. Experience has shown that researchers involved have appreciated the efforts, with many of them being encouraged to cooperate further. How this essential partnership has arisen and been nurtured will be explained below.

Economists Online: one international subject-based repository
The Economists Online service has become one of Nereus's pillars. What started out as a pilot project initiated by six partners (Erasmus University Rotterdam, ZBW – German National Library of Economics, London School of Economics, Maastricht University, Tilburg University, Université Libre de Bruxelles), a project which came to completion at the end of 2005, now is a service which includes 15 Nereus members. The pilot was led by Tilburg University with seed money from SURF, The Netherlands, although it was mainly financed by Nereus membership contributions.

Economists Online aims to increase the usability, accessibility and visibility of European economics research by digitizing, organizing, archiving and disseminating the complete academic output of some of Europe's leading economists, increasing access to full text. EO has done this by building an integrated international open access showcase of some of Europe's top economics researchers from some of its top research institutions based on institutional repository content. Local repositories are harvested by the Nereus Economists Online OAI repository, which then generates a number of added value services for the economist, which in turn stimulates the depositing of material into the IR by the economist.

The EO pilot contains material from six institutions, from four countries, with material in four languages (English, French, Dutch and German), although English is the lingua franca of economics. Close to seventy leading authors have been involved in the project, now with nearly 7,000 bibliographic references searchable, almost 40 per cent of which are available in open access, full text form (1,675 journal articles, 752 working and discussion papers, 269 chapters, 56 reports, 38 books and 21 conference proceedings and 53 other publications to date). User studies have shown that economists want access to more chapters and conference proceedings online, in addition to journal articles, and the project is starting to make a contribution to this need. These statistics are on the increase due to the enthusiasm of participating researchers even after the end of the pilot, a pilot that has now become a Nereus service. Full text coverage varies across countries and institutions, based on national copyright law, local IR policies, or self-archiving histories for example.

Several EO pilot partners have seen this as an opportunity to set up their IRs, doing so in collaboration with other experienced colleagues, other new Nereus members are following suit as in Leuven, Paris, and Dublin. All partners are therefore at different stages of IR development and rates of content acquisition. However, this initiative

being endorsed by Nereus partner faculties is probably the project's most important critical success factor; it is only a matter of time before IRs and their services really take off as long as the researcher's needs are the key starting point.

The researcher's needs and uncertainties
"The library is making my scientific life easier and more relaxed" is a quote by a leading economist in response to Economists Online. Researchers encouraged to participate in repository projects have often been known to say that the opposite is true.

IRs will not obtain the quality critical mass of content we strive for unless we build services that are of a direct benefit to the information provider. The top-down approach is clearly one way of getting results; however, for a dynamic and content-rich service, researchers need to be stimulated to contribute. Services need to be developed with the knowledge of the work processes and needs of the user in mind; this is what Economists Online has striven to do.

As a start, Nereus information specialists carried out user studies in 2005 into the information needs of a number of its institutions, and approached their faculties and authors regarding EO. Feedback given in these instances was utilized for the project; the goals were developed, as were its products. This was the start of a number of advocacy initiatives. At the beginning of the project, a short project leaflet was developed, addressing some of the researchers' interests such as visibility, showcasing and access to quality content, or archiving. Other PR materials were developed, e.g. a brochure, poster and website (all designed by a professional graphic designer), which directly spoke to the researcher to some of his/her direct concerns during the work process, offering concrete products in answer to some of those needs: What is in it for me? What is the added value? Why should I contribute? How we can support you. These lines were addressed in material and discussions, as were issues concerning copyright in individual talks with faculty and contributing scientists at scientific conferences. In some cases, a copyright overview document showing the policies of key publishers was presented to authors to inform them of which publishers allow or do not allow self-archiving before contributing their electronic files.

Also, as a result of feedback given, and discussions in the library community, a number of scenarios were drawn up for arguments when in discussion with researchers. These included for example: permanent access to your own research material "24 × 7" wherever you are located; a reference point containing your life's work for colleagues, faculty, the press; access to material to colleagues with little access to library online resources be they in smaller research institutions or in under-developed countries; assistance in updating and verifying your publications for your curriculum vitae; more visibility on the web in areas of importance to you. With these arguments, the researcher is beginning to see opportunities where libraries can step in an area that is outside of the scope of any individual publisher.

In order to promote the project, and to increase online open access content, researchers were also presented with interim project results, with statistics on how much full text was online with percentages in some case, also identifying gaps in online access to their content, and exploring how that content could also be made available. In many cases, this stimulated authors to contribute more content to the project. Researchers were also asked at the end of the project what added value they saw in contributing to the project to which they replied: the bibliographic reference lists with

new full text links as a good place for referral, further visibility and easier access to research results, as well as long-term archiving in some cases. This agreeably mirrored the original aims of the project.

We have also used scientific conferences to discuss EO with researchers and their dissemination and information retrieval practices, e.g. the European Economics Association in Amsterdam in 2005, to raise awareness of the project and what libraries are presently doing. Such venues are important to talk to the research community and to verify that present goals are still relevant.

Discussions with researchers identify needs, but also fears and apprehensions, which need to be well addressed before building the trusting relationships between library and client. Fears expressed included the lack of time, i.e. time is research, the importance therefore of keeping administrative tasks to a minimum. Concerns regarding copyright were responded to with information on the libraries' policies, the rights that authors have and the opportunities open to them in the future. Once entering into the project, authors found it important to know that the structural support was there to maintain the initiative in the future to ensure that current content would also be added to the system; all libraries are committed to maintaining and indeed expanding on the service in the future.

As a result of having a meaningful product (to be demonstrated below) and with arguments that meet user needs and dispel fears, content is increasing in the EO pilot IRs. New partners from other countries are now joining where faculty heads have been approached with the same advocacy messages, resulting in enthusiasm to participate. In some cases it is the faculties who are putting up the annual Nereus membership fees. These are signs that Economists Online is achieving its aim.

Translating needs into services: the Economists Online showcase, search service, publication lists and repository – A demonstration

The following screenshots demonstrate the Economists Online service at work (Figure 1).

The EO service main page publicizes the current EO partners linking to the index of some of their top names in economics research via an institutional link; profiling the institution first (Figure 2).

This page shows the names of an institution's leading economists, in this case a selection from the London School of Economics (LSE). Economists Online names were mainly selected by faculty. Their names, and affiliations are indicated, and further information on the individuals can be retrieved from either the EO generated "publication list" link (see Figure 3), or by clicking on "more" which takes you to the personal home page of the author.

This shows a publication list which has been generated by the EO repository via xml; using the bibliographic format of American Psychological Association (APA). Publications are sorted by type. A resource type index at the top of the page allows one to jump to the type of publication of one's choice. "More"-links generally bring one to the jump-off page of the repository record (Figure 4).

A simple search toolbar at the top of the EO service page allows one to search the EO repository metadata, fed by all Nereus repositories. "All fields" is the default search although title words, author and year can also be searched for. A full text search service is planned for EO in the future (Figure 5).

Figure 3.
An EO author publication
list

Figure 4.
Search the Economists
Online service

Figure 5.
Search results from the
Nereus EO repository
(all repositories)

Search results are sorted by date and then by title. Links take you through to the metadata record, and for access to full text material, then linking on to the jump-off page of the repository or directly to the document (Figure 6).

This jump-off page comes from Tilburg University; it shows links to the publisher website and article as well as to the open access post-print. Figures 7 and 8 demonstrate the publications available for download.

Figure 6.
Local partner jump-off
page (varies per partner)

Figure 7.
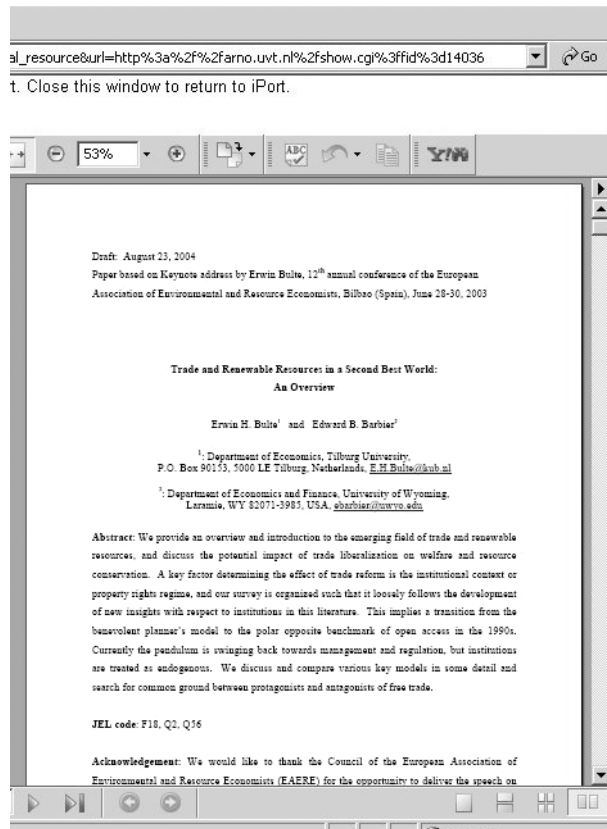Link through to the
publisher version of the
article

Figure 8.
Link through to the open
access, free post-print
version of the article

Conclusion

IRs can prove to be a significant catalyst for bringing libraries closer to their researchers and to the research process, provided they are designed with the needs of the researcher in mind. Economists Online is a case in point, in which partnerships between management, researcher and the international library/faculty community have brought benefits to all. It is a model that is complimentary to the services that publishers provide. EO focuses on developing services of direct value to the author, the content-provider, e.g. providing new full text content online, digitizing older material, creating automated publication lists, metadata quality-control, more focused dissemination of content. Once the partnerships between libraries and researchers have been enhanced, through strong products and effective advocacy, experience has shown that the interests of the institution (e.g. increased visibility of research) and its library (high quality IR content) converge. Economists Online could well be a model for others to follow in the future.

For more information, please go to the Nereus website[5] or to the Economists Online site[3].

Notes

1. www.eeassoc.org/default.asp?AId $=$ 37
2. www.nereus4economics.info/rolling_plan.html
3. www.nereus4economics.info/econline.html
4. www.lse.ac.uk/versions
5. www.nereus4economics.info

Reference

Coupé, T. (2003), "Revealed performances: worldwide rankings of economists and economics
    departments, 1990-2000", *Journal of the European Economic Association*, Vol. 1 No. 6,
    pp. 1309-45.

Corresponding author
Vanessa Proudman can be contacted at: v.m.proudman@uvt.nl