TECHNICAL
REPORT

Addressing the uncertain future of preserving the past

Towards a robust strategy for digital archiving and preservation

Stijn Hoorens, Jeff Rothenberg,
Constantijn van Oranje, Martijn van der Mandele,
Ruth Levitt

RAND EUROPE

TECHNICAL
REPORT

Addressing the uncertain future of preserving the past

Towards a robust strategy for digital archiving and preservation

Stijn Hoorens, Jeff Rothenberg,
Constantijn van Orange, Martijn van der Mandele,
Ruth Levitt

Prepared for the Koninklijke Bibliotheek

RAND EUROPE

The research described in this report was prepared for the Koninklijke Bibliotheek.

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND® is a registered trademark.

© Copyright 2007 RAND Corporation
All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from RAND.

Published 2007 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom
RAND URL: http://www.rand.org/
RAND Europe URL: http://www.rand.org/randeurope
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

This document examines key determinants of the sustainable digital preservation of scholarly records, with specific reference to developing a robust approach to the archiving of such records at the Koninklijke Bibliotheek, in The Netherlands, which commissioned and funded this study.

The purpose of the document is to analyse the Koninklijke Bibliotheek's e-Depot strategy in the context of wider developments in the archiving and publishing environment, develop scenarios for future framework conditions and highlight a range of strategic options to sustain the Koninklijke Bibliotheek's successful record in digital preservation. Storing and curating authentic academic literature and making it accessible for the long term has been a time-honoured task of the Koninklijke Bibliotheek, as of other national libraries. By guarding existing knowledge and facilitating its use to produce new insights, national and university libraries have formed an integral part of the research ecosystem, complementing the roles of other stakeholders such as researchers, publishers and funders. However, recently the digital revolution has modified fundamentally the way that research results are circulated, reviewed, accessed and preserved. Hitherto established models of market dynamics and stewardship need to be rethought and part of the responsibilities of national libraries redefined.

The Koninklijke Bibliotheek is among the pioneers of digital archiving and has demonstrated a degree of concern with long-term strategy that is rare among the stakeholders concerned. Such long-term strategy requires a sound evidence base, but it also must acknowledge the need to respond flexibly to a variety of possible future developments and events. Based on desk research and international expert interviews, this report was produced to answer both these requirements.

The report will be of interest to staff at the Koninklijke Bibliotheek, as well as decision-makers at other national and university libraries faced with the challenges of digital archiving and preservation. It will also be relevant to other stakeholders in the distribution of scholarly knowledge, including traditional and open-access publishers, operators of dissemination portals, researchers, learned societies, journal editors, funders, developers of preservation technologies and governments.

RAND Europe is an independent private, not-for-profit, research institution which helps to improve policy and decision-making through research and analysis. RAND Europe is an independently-chartered European unit of the worldwide operating think tank, the RAND

Corporation.[1] For more information about RAND Europe or this document, please contact:

Stijn Hoorens
RAND Europe
Westbrook Centre
Milton Road
Cambridge CB4 1YG
UK
Tel: +44 1223 353329
email: hoorens@rand.org

Jeff Rothenberg
RAND
1776 Main Street
PO Box 2138
Santa Monica, CA 90407
USA
Tel: +1-310-393-0411
email: jeff@rand.org

[1] For more information about the RAND Corporation and RAND Europe, please see: http://www.rand.org and http://www.randeurope.org.

Contents

Preface iii
List of figures, tables and boxes vii
Executive summary ix
Acknowledgements xix

CHAPTER 1 **Introduction 1**
1.1 Why digital preservation is important 1
1.2 Scope, objectives and approach 1
1.3 Structure of this report 4

CHAPTER 2 **A brief history of preserving the digital past 5**
2.1 A new era for academic publishing 5
2.2 Concerns about perpetual access 6
2.3 Digital archiving and preservation 7
2.4 KB and the e-Depot 9
2.5 Defining some terms 11

CHAPTER 3 **Scholarly dissemination and publishing: a complex and dynamic environment 15**
3.1 The outlook of scholarly dissemination and publishing: key figures and trends 15
3.2 Stakeholders' views on scholarly dissemination and publishing 22
3.3 Trends and uncertainties in scholarly dissemination and publishing 26

CHAPTER 4 **Digital archiving and preservation: an area under construction 33**
4.1 The outlook of digital preservation: key figures 33
4.2 Stakeholders' views on digital preservation 35
4.3 Assessment of three preservation models 37
4.4 Costs of archiving and preservation 44
4.5 Trends and uncertainties in preservation 46

CHAPTER 5 **The uncertain future of preserving the past 53**
5.1 Scenario development 53
5.2 Driving forces of digital preservation 55

5.3 Possible future scenarios for dissemination and preservation 59
5.4 Conclusion 62

CHAPTER 6 **Strategy implications for KB 63**
6.1 The continuing justification of KB's activities 63
6.2 Awareness and trust in the stakeholder community 65
6.3 Justify KB's national and international remit 67
6.4 Need for a sustainable funding model 69
6.5 More transparent access rights policy 71
6.6 Coordination between initiatives and peers 73
6.7 Specify scope of preservation 74
6.8 Robust planning of an e-Depot strategy 76
6.9 Conclusion 77

References 81

Appendices 87
Appendix A: Summary of key assumptions underpinning the e-Depot strategy 89
Appendix B: List of interviewees 91
Appendix C: The technological basis for digital preservation 95
Appendix D: 12 e-journal archives compared 103
Appendix E: Stakeholders' views and positions on scholarly communication and publishing 105
Appendix F: Stakeholders' views and positions on preservation 113

List of figures, tables and boxes

Figure 1. Higher education expenditure on research and development by field of science in selected countries 6

Figure 2. Global market shares in STM publishing, 2003 16

Figure 3. STM publishing market breakdown by delivery format, 2006 17

Figure 4. Growth of the number of academic journals over time (historic and projected) 18

Figure 5. Average journal prices from for-profit and not-for-profit publishers 19

Figure 6. Average article prices from for-profit and not-for-profit publishers 19

Figure 7. Usage statistics by subject area 21

Figure 8. Cited half-life distribution of the top 500 journals in *JCR Science Edition* 22

Figure 9. The building blocks of scenarios 54

Figure 10. Possible compatibility of different functions of digital archives 58

Figure 11. Scenario framework for the future of digital archiving and preservation 60

Table 1. Life expectancy of digital media and the expected time until obsolete 8

Table 2. Proportion of readings by age of scholarly articles by university scientists 1993–1998 20

Table 3. Summary of 12 included digital archiving initiatives 34

Table 4. Objects ingested in e-Depot as of 1 August 2007 39

Table 5. Estimated cost breakdown of KB's e-Depot, distinguishing the international and national e-Depot 45

Table 6. Different rationales for digital archives 56

Table 7. Archiving strategy: what type of archiving strategy do the e-journal archives use? 102

Table 8. Documentation type: what kind of written documentation do e-journal archives (plan to) have that explicitly refers to e-journal archiving? 103

Table 9. Trigger event: trigger events that spark changes in access for the authorised community. 103

Table 10. Archiving activity: do the e-journal archives have any relationships with other archiving organisations involving the following activities? 104

Box 1. Trigger events prompting the need for perpetual access 7

Box 2. Clauses 6–8 of the 'Brussels Declaration' 107

Executive summary

As part of the public responsibility of the national library of The Netherlands to archive publications with a Dutch imprint - general publications as well as scholarly output - the Koninklijke Bibliotheek (KB) has undertaken to develop a digital archive, the national e-Depot. The 'e-Depot' system has been devised and implemented to maintain and preserve the delivered content for perpetual access. In addition to items having Dutch imprint, the KB archives international publications in the areas of science, technology and medicine (STM). Because the progress of scholarly knowledge feeds on the scholarship of the past, this is a task of great significance. Agreements have been made with the major international STM publishers, and with this global application of the e-Depot, KB aims to extend its public deposit function for electronic publications to the international level. In so doing, KB intends to become part of a global 'safe places network' consisting of a limited number of digital repositories for international electronic publications.

KB has asked RAND Europe to assess the premises on which the 2006–2009 strategy for its international e-Depot for STM publications was based. Based on this review, RAND Europe has come to a number of conclusions regarding the future of archiving in STM publications and the role that KB can play. In many cases, these conclusions support existing plans and policies of the KB, so they should not be construed as criticizing these plans and policies but rather as reinforcing and substantiating them.

Digital archiving and preservation is an important but intractable issue

In the wake of the digital revolution, stewardship of learned publications has acquired new opportunities as well as highly complex dimensions. These include fundamental shifts in the relationships between libraries, publishers and researchers. In their traditional role as custodians of society's accumulated knowledge, librarians face new challenges with regard to the access and preservation of digital information.

Scientific and scholarly disciplines – as well as society as a whole – have a crucial need to preserve an authentic scholarly record for intellectual, scientific, pragmatic, legal, historical and ethical purposes. Techniques for preserving traditional artefacts in the print environment are inadequate for preserving many kinds of digital artefacts. If the present and future scholarly record is to be saved in its full and authentic form, it is essential to find ways of preserving digital scholarly material.

One aspect of digital preservation that is rarely discussed is the problem that most proposed techniques, including migration, are based on the repeated conversion of digital

objects into successive new formats over time. Aside from the fact that such approaches do not even attempt to preserve a digital object in its original form, their repeated conversion leads to the inevitable cumulative corruption and degradation of each digital object as it is force-fit into the Procrustean bed of each successive digital format.

Moreover, despite the fact that most current scholarly publication still produces traditional "page-image" artefacts (i.e. static objects that are the digital equivalent of printed material), a crucial and growing category of scholarly output – particularly in STM – consists of "inherently digital" artefacts that must be interpreted by software that is "executed" (i.e. run on a digital computer) in order to be rendered into perceptible form. Without such rendering, these inherently digital objects cannot be viewed or used at all and may in a fundamental sense not even be said to exist, since their content may be generated or constructed only as they are interpreted by software. In some cases, these inherently digital objects are themselves executable programs. In general, inherently digital objects are characterised by dynamism, interactivity and complex behaviour that has no analogy to page-image objects and cannot be printed or preserved by static techniques. Because of the rapid evolution of information technology, the software necessary to render such inherently digital objects – and the computer hardware on which this software runs – can become obsolete in just a few years. If preservation methods cannot preserve their full range of behaviour, the future scholarly record will bear only static, snapshot representations of the first generation of these inherently digital objects, which are likely to become increasingly numerous and important to scientists and scholars over time. Such static representations will not constitute an authentic scholarly record of STM.

It is therefore important to distinguish between "static preservation" techniques (such as migration), which may be adequate for page-image artefacts, and "behaviour preservation" techniques (such as emulation), which are necessary to preserve inherently digital artefacts. As inherently digital artefacts become increasingly common and important in STM publication, the need for behaviour preservation is likely to become increasingly crucial. Evidence to date is that behaviour preservation techniques such as emulation are likely to be less expensive (as well as less error-prone) than static techniques – even for preserving static, page-image objects. Nevertheless, behavioural techniques like emulation are still unfamiliar (and even mysterious) to much of the scholarly community and are therefore viewed with scepticism. In addition, continued research and development are needed to refine behaviour preservation techniques and improve their usability.

KB's international e-Depot strategy is built on three main principles

KB was among the first to recognise the issues surrounding digital preservation. Since off-the-shelf systems for digital archiving and preservation have not been available, KB investigated the possibility of developing and implementing the e-Depot system for digital publications. The principles upon which the KB's strategy is based can be summarised as follows.

1. *Archiving and preservation of digital objects.* As scholarly output is moving toward the exclusive use of electronic form, a digital archive is needed for KB to continue fulfilling its deposit task as a national library. Archiving and preservation of digital

objects is fundamentally different from archiving and preservation of print objects. Archiving and preserving electronic publications for the long-term, in order to safeguard future access to their original intellectual content, requires a substantial investment in infrastructure, equipment, skills and expertise.

2. *International deposit function.* Since the concept of imprint (location of publication) is no longer valid for digital publications, KB extends its national deposit function to the international level. In so doing, KB offers to become part of a global 'Safe Places Network', consisting of a number of digital repositories for international electronic publications. Because of the required scale of investment in equipment, skills and expertise, as well as a consequence of publishers' archiving policies, it is expected that there will be a limited number of such 'safe places'.
3. *Perpetual access.* KB acknowledges research libraries' concern about the threat of permanent loss of electronic journals and disrupted access to journals for a protracted period following a trigger event, such as a publisher going out of business or a library cancelling a journal subscription (see Section 2.2). An e-Depot would provide a way to manage this threat. Following a trigger event, e-Depot would provide affected libraries with either temporary or permanent access to a specific set of serials and volumes in its archive.

In assessing KB's strategy and the wider context of scholarly dissemination, STM publishing, digital archiving and preservation, we have taken different timescale perspectives. Consequently, we have distinguished short or medium-term conclusions and recommendations from long-term conclusions and recommendations.

Short-term conclusions and recommendations

KB seems well positioned to play an important international role in the digital archiving of STM publications

National libraries and archives have long taken a central role in preserving national and collective heritage, as custodians of the 'collective memory'. Therefore, they are well positioned to play an important role in the international coalition for long-term preservation of digital scholarly output. Because information hubs, search engines and publishers will not provide the guarantee of structured, perpetual access to entire collections, it is likely that this will have to remain a public service function. There are several possible candidates for the role of a reliable safe keeper who will ensure continued availability of electronic records even if established access mechanisms are disrupted by a 'trigger event'. Among these candidates, KB commands a number of distinguishing merits. In particular, preservation is already one of its core functions: it possesses recognised expertise in the area and pursues no commercial goals that may conflict with careful archiving. As early as the mid-1990s, KB acknowledged the resulting challenges to the preservation and archiving of digital publications. KB has invested in multi-pronged research and development (R&D) in digital archiving and preservation to guarantee perpetual access to authentic scholarly material. These ongoing R&D efforts have brought KB international recognition for its expertise in this area in and beyond expert circles. Since providers of related services such as Portico, LOCKSS (Lots of Copies Keep Stuff Safe) and CLOCKSS (Controlled LOCKSS) are so new, it is difficult to assess their track

record and consequently their credibility as reliable preservation institutions. As a national library, KB is seen as having a longer-term perspective, no financial motives and an orientation toward scholarly access. KB also benefits from a neutral position, being a governmental institution in a small and relatively neutral country.

In order to play its international role, KB should increase awareness and trust among the global stakeholder community

Among experts, KB is well known for its prominent position within the state of knowledge on digital archiving. However, wider and international stakeholder audiences, for example in Asia and North America, are often unaware of KB's expertise in the area. In order to develop its role as a provider of specialised supranational and supra-institutional services and to forge productive partnerships with others, it appears strongly advisable that KB promote its expertise among the many organisations which have yet to devise a robust approach to digital archiving, including university libraries. The latter form a crucial stakeholder group in the area of digital archiving, and are themselves in the middle of finding new sustainable roles in the digital era. Many of these university libraries – some of which worry that KB may pre-empt the function as a provider of research information services which traditionally have been an important part of their brief – see themselves as having been ignored by KB. Thus it is important for KB to: spell out what will bind it firmly to provide 'access insurance' to such libraries, in the way that insurance businesses are under legal and moral obligation to fee-paying customers; seek membership in an international consortium to preclude national bias in preservation; and provide precise details of how records would become available in the wake of a trigger event. KB's mission and approach to STM archiving should be made widely known in the international community of scholars, scholarly societies, research libraries, publishers, information hubs, digitisers and archives.

KB should seek to develop a sustainable funding model

The development of the e-Depot and the pre-eminent position that KB currently holds in digital preservation has been very dependent on the active support of the Dutch government. Continuation of this support is indispensable to realise KB's plans and create a business that can continue to carry its own (economic) load in perpetuity.

At the same time, this dependence on government funding raises questions in the scholarly community as to the independence and sustainability of KB. In addition, the greater demands on archiving that come with international scope may require a new funding model, which might include additional resources. Therefore, KB should establish the market value of its services and provide a detailed overview of the costs of its e-Depot operations. Additionally, it should consider developing independent sources of funding for its e-Depot activities. There are indications that the large publishers are willing to economically underwrite the work of KB, and other stakeholders in the research library community acknowledge that the services provided by KB may be worth their supporting. Yet the library and academic community is not a very commercial one, and the lack of awareness that exists in some quarters about KB's activities means that additional work

may have to be done by KB if it is to develop this potential source of economic support. In addition, it might be possible for KB to offer its research capacity through consultancy services or research projects, which might broaden its funding base. At the same time, KB should assure itself and its stakeholders of the continuing support by the Dutch Government of its activities and archives. The fact that KB is an acknowledged innovator in its field is valuable in its own right, and KB should articulate the benefits of being an international centre of excellence – a concept which may fit into Dutch governmental policy to promote a knowledge-based economy – in order to help ensure the required level of public funding. In any case, the possible development of a broader base of economic support must not be allowed to undermine KB's independence and neutrality, which underlie much of its credibility and reputation.

KB needs to clarify access to e-Depot content prior to, and following, a trigger event

Providing access to its collections is a core function of a (national) library. However, in the digital environment this conflicts with the interest of the publishing industry, which retains copyright on the published content. There still are substantial disagreements as to what trigger events should initiate such access and what the consequences are. Not all publishers are comfortable with preservation models that grant open access (before copyright has expired) in case of a trigger event, as they will lose control over content. While the ideal situation would be for a safe place to provide access mechanisms that are limited to those who have license rights to digital content, current solutions do not meet those needs. To address at least some of this concern, KB should improve communication of its definitions and conditions of trigger events and clarify what services can be expected under which circumstances. These policies should be communicated to libraries and publishers in order to allow them to anticipate the outcomes of such events.

Additionally, KB should evaluate the e-Depot access regime prior to a trigger event. Currently, e-Depot can be considered as a 'dim archive' that is neither dark (providing no access) nor light (providing unlimited online access) but is accessible only to users on-site at KB's premises. While this policy is considered by KB to be a non-financial compensation for its free service to publishers, it is met with scepticism by several stakeholder groups. This ambiguous compensation scheme compromises the transparency of the cost structure of preservation services, which some stakeholders believe should be set by the market value of such services and the costs related to managing access in case of trigger events. Policies concerning access to such content would need to be addressed as a separate issue and should preferably be based on regular licensing agreements with publishers.

KB's vision of a safe places network is endorsed, but its development lacks a discussion platform and leadership

An international perspective, economies of scale, mutual auditing, diversification of risk and replication of records are some of the key advantages of a safe places model to guard digital records across Europe and the rest of the world. In signalling its interest in becoming part of a safe places network, KB has recognised the immediate as well as wider

potential benefits of this model. Stakeholders, including publishers, generally embrace this vision, but there is the perception of lack of leadership in this development and the absence of an effective platform for discussion between the main (national) preservation libraries.

The existing initiatives involved in preservation of electronic journals (e-journals) all have distinct mandates, funding sources, business models, temporal outlooks, preservation strategies, arrangements with publishers and relationships with scholars. It would be to the mutual benefit of these initiatives to establish relationships with each other. Also, it would be natural to share costs and technology, exchange best practices and cooperate in agreeing on common standards. Because KB has a track record and credibility as a reliable preservation institution, it is important that any such relationship with other digital archives avoid compromising any of this credibility – and indeed enhance it, if at all possible. This suggests that KB should guard against diluting any of its key advantages in such a relationship, i.e. its financial independence from publishers, its multi-pronged, long-term preservation perspective and its orientation toward scholarly access.

Medium-term conclusions and recommendations

KB will need to continuously monitor emerging trends in scholarly dissemination and publishing

In developing its approach to the preservation of digital records, KB is undertaking a highly-challenging endeavour that will be characterised by constant tension between consistency and change, as well as between conception and implementation. To make effective preservation possible, it will be indispensable to set down definitions and select the contemporary approach to archiving that provides the best available match with requirements for authenticity and durability.

Slowly but surely, the use of non-page image objects is intensifying. Such objects include:

- dynamic webpages;
- animations;
- video and other multimedia;
- databases;
- geographic information systems;
- models and simulations;
- finer-grained units of information and embedded objects;
- virtual compound objects and inherently digital objects, including script-generated webpages and executable models; and
- visualisations and programs of all kinds.

Although these new types of objects do not yet constitute a dramatic percentage of the scholarly record, they seem likely to become increasingly numerous and important over the next 10 to 20 years, if not well before then. KB's multi-pronged approach to preservation – particularly its use of emulation – is well suited to preserving such objects.

Nevertheless, in order to achieve its objectives as a national institution as well as a leading player in international efforts to safeguard learned knowledge, it will be essential for KB to monitor emerging technological developments relevant to electronic archiving. Timely and continuous consideration of such developments will allow KB to adapt its preservation strategy and coordinate any necessary changes with its partners. In particular, KB should continue its research and development of behaviour preservation techniques that can cope with multiple formats, many of which are likely to include inherently digital content.

Continuously review the boundaries of the records of science, but strive for completeness

KB will be challenged – first and foremost in its public service role as the archive for national STM and other published digital output – to identify what elements of the scholarly record it should target for preservation. In fact, the selection of material worthy of archiving could become more time-consuming than archiving everything. Obviously, the choice to include other kinds of content than STM publications involves the much broader societal question of what needs to be kept for future use: what is the future of archiving in a time when information is being generated everywhere and at unprecedented speed? In order to archive and preserve the records of science, it is vital to continuously review what constitutes these records. In this regard, KB's ongoing examination of preservation techniques for institutional publication, self-publication and web publication should be continued and expanded as necessary.

The safe places model offers a suitable solution for capturing such a wide variety of content from such a wide range of sources for archiving. Publications from smaller publishers and less prestigious journals would benefit from this model. For KB (and its peers) it is important to consider how to encourage relatively small publishers, whose content is more susceptible to becoming inaccessible, to participate in archiving schemes. One way of facilitating such participation would be to make depositing material as easy as possible, so that compliance costs are minimised. At the same time, effective guarantees must be in place to protect publishers' copyright, thus creating an environment of trust.

Long-term conclusions and recommendations

Three key assumptions underlying KB's strategy are critical in the uncertain long-term future of preservation

The future of the organisation of digital archiving and preservation is uncertain, and off-the-shelf technical solutions are not (yet) available. Because developments in publishing, scholarly dissemination, information services and technology are rapid and dynamic, developing an e-Depot strategy will need to anticipate these uncertainties and prepare for future trends that are foreseeable. We feel that KB's strategic choices involve three critical assumptions about the uncertain future of digital preservation. Without implying that KB is unaware of these assumptions or is not taking steps to address them, we nevertheless feel that they deserve to be made explicit:

1. KB aims to sign archiving agreements with the 20 to 25 largest publishing companies which produce almost 90 percent of the world's electronic STM literature. The assumption underlying this objective is that large traditional publishers will continue

to be the main providers of scholarly content, although KB is also pursuing other sources of STM literature, including institutional publication, self-publication and web publication.

2. KB believes that it is essential to preserve the authentic content, format and behaviour of digital objects, because it is impossible to predict which attributes of an object may be important in the future. Consequently, KB considers preservation of the original object (or something very close to the original) to be essential, along with the ability to preserve the original behaviour of the object.
3. KB assumes that government funding for its preservation initiative will continue into the long-term future.

Although these assumptions about long-term developments may well be warranted, they are nevertheless critical because they are uncertain and outside KB's sphere of direct influence. Looking at the future of digital archiving and preservation, we identify two main dimensions of uncertainty.

1. The future outlook of scholarly dissemination and STM publishing faces a period of uncertainty in which the role of an international repository is unclear

Over the years, traditional journal publications have been the prime channel of scholarly dissemination. This has been a remarkably robust and effective means of dissemination. The publishing market has been dominated by a number of large companies which have been in a position to maintain networks of scholars, support peer review and distribute the resulting material to libraries and individuals. More recently, there has been a gradual shift towards a more diverse portfolio of dissemination channels. Examples of these alternative channels include models based on traditional publications, such as open-access publishing, but also more informal methods of dissemination, such online portals, weblogs (blogs) and wikis. In parallel, information hubs, particularly those that include search engines such as Google and Yahoo! offer technologically sophisticated routes to information. These services facilitate a 'demand pull' process for scholarly dissemination, rather than a 'supply push' from a publisher's periodical table of contents.

While the impact of these alternative channels is currently very limited and may vary by field, a trend break is not an infeasible scenario: the traditional publishing model may or may not continue to dominate the STM sector. It remains unclear how these new scientific information review and dissemination channels – be they of the 'push' or 'pull' type – will manage their preservation and archiving needs. In organising digital archiving and preservation, a current focus on traditional publishers is understandable, particularly because at present there are few alternatives. However, this may not necessarily be a robust strategy in an uncertain future.

2. Stakeholders' priority and needs for archiving and preservation now and into the future are yet unclear

Based on traditional preservation needs, it seems likely that many stakeholders will demand the authentic preservation of original artefacts, along with all of their original behaviour.

Demand for such authenticity is found among traditional scholars, as well as curators of cultural heritage and governmental and societal institutions seeking historical accountability. However, the current demand for the behaviour preservation of digital objects is quite low in the scholarly community. This appears to be due to several factors. First, the relative paucity of inherently digital artefacts in current scholarly production has so far kept awareness of the need for behaviour preservation to a minimum. In addition, the scholarly community has not yet experienced the cumulative corruption that is the inevitable result of repeated conversion of digital artefacts into successive new formats. Finally, a lack of understanding in the scholarly community of techniques such as emulation has led to scepticism about behaviour preservation. If the demand for behaviour preservation remains low, it may be difficult to convince the scholarly community of the value of mechanisms (such as emulation) that offer authentic preservation of the behaviour of inherently digital objects, even though the evidence to date is that these techniques are less costly – even for preserving page-image objects – than other, less robust alternatives. It is as yet unclear whether or when demand for the behaviour preservation of digital scholarly originals will emerge, although it appears likely to do so, as inherently digital artefacts come to comprise an increasing proportion of the scholarly record.

Initiate a process of robust strategic planning using possible scenarios for the future of archiving and preservation

Using the two dimensions introduced above, we can develop four possible scenarios for the future of digital archiving and preservation, each representing unlikely, but not inconceivable pictures of the long-term (10 to 20-year) future.

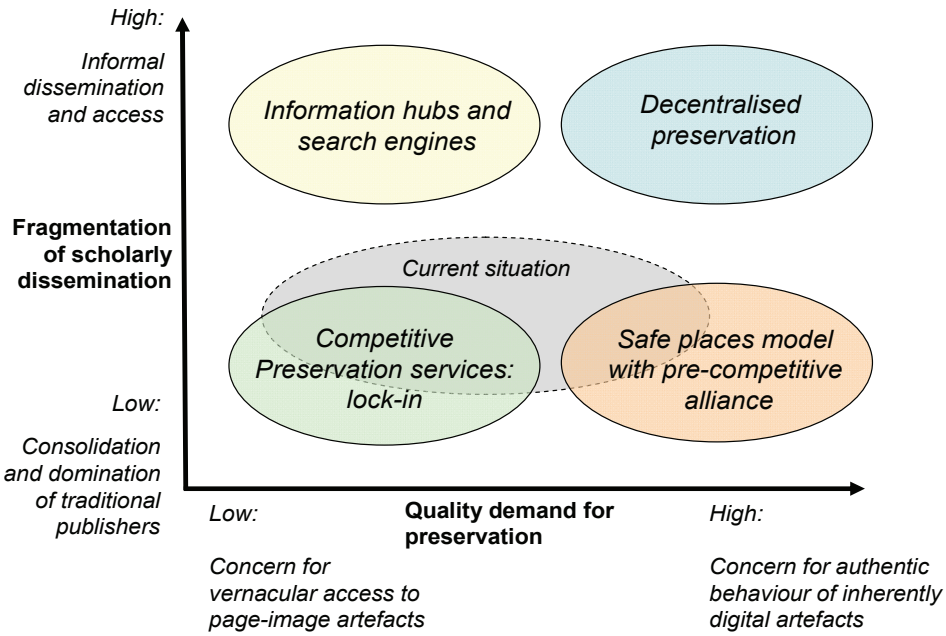

Figure A: Scenario framework for the future of digital archiving and preservation

The grey oval represents the situation in the wider stakeholder context, with still relatively little fragmentation in scholarly dissemination and relatively low priority for behaviour preservation of inherently digital artefacts.

To address the exogenous factors that will affect the future of KB's strategy in a fast-changing global setting, KB should consider testing its strategic options against these scenarios. Scenarios are not predictions, but end-of-spectrum examples that are useful for illustrating the range of plausible futures – in this case, over the next 10 to 20 years. They provide a framework to consider uncertainties, key drivers and policy levers that will determine the context in which KB has to operate in the near, mid- and long-term future.

We recommend that KB conduct a process of robust strategic planning through testing its strategic assumptions against a set of such scenarios, in order to help address the uncertainties in the market for long-term digital preservation. These scenarios can be used to communicate with internal and external stakeholders and actively engage them in KB's planning process, in order to address uncertainties and develop a shared vision of how to address them.

Acknowledgements

This report would not have been possible without the support of numerous individuals. The RAND team would like to express its appreciation to the sponsor, Koninklijke Bibliotheek. In particular, we are grateful to Wim van Drimmelen, Hans Jansen, Els van Eijck van Heslinga, Johan F. Steenbakkers, Ingrid Dillo, Hilde van Wijngaarden and Erik Oltmans.

During the preparation of this report, we interviewed numerous stakeholders and experts including, but not limited to, those in the publishing world, library community and government. These individuals gave generously of their time to provide their views on the subject of this report. Therefore, we wish to thank the following people for their willingness to speak to us and for their valuable contributions (in alphabetical order): Reinhard Altenhöner, Martha Anderson, Paul Ayris, Kurt de Belder, Cornelis van Bochove, Pieter Bolman, Richard Boulderstone, Herman J. Bruggink, Dan Burnstone, Matthew Cockerill, Paul Courant, Robin Dale, Raymond van Diessen, Joachim Driessen, Eileen Fenton, Carlos Ferreira Morais-Pires, Dale Flecker, Nick Fowler, Amy Friedlander, Joost Geise, Steven Hall, Maria Heijne, Peter Hendriks, Karen Hunter, Wim Jansen, Kathleen Keane, Alice Keller, Robert Kiley, Reinier A.M. van Langen, Deanna Marcum, Ronald Milne, Frank L.A. van Oudenhove, David Prosser, Xu Qiang, Miao Qihao, Victoria A. Reich, David Rosenthal, Mary Sauer-Games, Ute Schwens, Abby Smith, Herman P. Spruijt, Dave Thomson, Frans Visscher, Donald J. Waters, and Andy Williams. (The affiliations of these individuals are given in Appendix B.)

We are also grateful to Dr Chris van Stolk (RAND Europe) and Dr Lorenzo Valeri (Accenture), who provided useful and insightful comments during the quality assurance process. Finally, we wish to thank Lynne Saylor, Josh Levine and Lisa Cordaro for their help during the publication process.

CHAPTER 1 Introduction

1.1 Why digital preservation is important

The digital revolution has offered a myriad of new opportunities across nearly all sectors. The daily activities of librarians and archivists have benefited enormously from machine processing, database technology, increased storage capacity and electronic content delivery, to name just a few obvious examples. However, in their traditional role as custodians of society's accumulated knowledge, librarians face new challenges in accessing and preserving digital information.

Scientific and scholarly disciplines – as well as society as a whole – have a crucial need to preserve an authentic scholarly record for intellectual, scientific, pragmatic, legal, historical and ethical purposes. Techniques for preserving traditional page image artefacts in the print environment (documents, photographs, drawings, etc.) are inadequate for preserving many kinds of digital artefacts. If the present and future scholarly record is to be saved in its full and authentic form, it is essential to find ways of preserving digital scholarly material.

Digital publications pose novel challenges for preservation, due to the short lifespan of most digital storage media: the potentially dynamic, multimedia, interactive behaviour of 'inherently digital' artefacts, the rapid obsolescence of the digital formats of such artefacts and the need to run appropriate software to interpret and render those formats into forms that humans can perceive and use. Because the formats of such artefacts – as well as the programs that interpret them and the computers on which those programs run – can become obsolete in just a few years, new methods must be developed to preserve their full range of behaviour. If this is not done, the future scholarly record will bear only static, snapshot representations of the first generation of these inherently digital objects, which are likely to become increasingly important to scientists and scholars over time.

1.2 Scope, objectives and approach

As part of the public responsibility of the national library of the Netherlands to archive scholarly output with a Dutch imprint, the Koninklijke Bibliotheek (KB) has undertaken to develop a digital archive for publications in the areas of science, technology and medicine (STM). Agreements have been made with the major international STM publishers, and the 'e-Depot' system has been devised and implemented to maintain and

preserve the delivered content for perpetual access. With this e-Depot, KB aims to extend its public deposit function for electronic publications to the international level. In so doing, KB intends to become part of a global 'safe places network' consisting of a limited number of digital repositories for international electronic publications (Oltmans and Lemmen 2006).

KB's vision of becoming one of the main nodes in the safe places network is untested. It is not yet clear how appropriate the above design principles and their underlying assumptions are.[2] The future of archiving and preservation is highly uncertain and off-the-shelf solutions are not yet available. Because developments in publishing, information services and technology are rapid and dynamic, developing an e-Depot strategy is akin to jumping on to a moving vehicle. A robust strategy will have to anticipate these uncertainties and prepare for future trends that are foreseeable. This report will explore the factors that may affect KB's vision (including varying degrees of uncertainty, some within and some outside KB's control), assess their relevance for KB and suggest ways forward.

In Appendix A, we summarise the main assumptions underlying the international e-Depot strategy. We will not assess individually the validity of these assumptions. Rather, we will show that there are a number of observations, trends and uncertainties which may have an impact on KB strategy. We have analysed three areas that we considered relevant for KB's strategy:

- *scholarly communication* – this includes all aspects of the shift in what is being produced by scholars and how they are disseminating, accessing and using it;
- *publishing* – this includes shifts in the publishing industry's composition, business models, temporal horizon (e.g. 'long-tail' value issues), etc. It also includes issues of copyright as seen from the publishers' perspective;
- *preservation* – this includes the technological, organisational, financial and regulatory or legal issues surrounding preservation, such as how it can be done, how it is affected by new forms of digital material, who should perform it, and how it can be paid for (overall preservation business models), etc.

In these areas, we have delineated historic trends, the current state of play, stakeholders' perspectives and future developments and uncertainties. Since the fields of scholarly communication and publishing are so closely intertwined and are difficult to isolate, we have chosen to discuss them together.

We have assessed the e-Depot strategy, taking both a short or medium-term perspective and a long-term perspective. For the short and medium-term perspective, we have conducted a stakeholder analysis and assessments of the e-Depot business model and two other key digital preservation initiatives. Using the findings from these assessments, we have derived conclusions and recommendations for the strategy of KB. For the long-term perspective, we have applied a simplified scenario planning approach, identifying the most uncertain developments which will have a high impact on the future of preservation. For more information on scenario planning, see Chapter 5. Having identified four possible

[2] For a summary of the main underlying principles of KB's international e-Depot strategy, see Appendix A.

scenarios for the future of digital preservation, we suggest how KB (and other organisations concerned with digital preservation) may anticipate these uncertain developments and plan a robust strategy in a highly uncertain environment.

To this end, this report attempts to address a number of key questions:

Background:

- What is digital preservation, how did it become an issue and how did KB get involved in it (Chapter 2)?

Scholarly dissemination and publishing:

- What is currently the outlook of the environments for scholarly communication and publishing (Section 3.1)?
- What are the needs and interests of the relevant stakeholders (Section 3.2)?
- What are current and future developments in scholarly communication and publishing, and which will have an important impact on the future of digital archiving and preservation (Section 3.3)?

Digital preservation and archiving:

- What is currently the outlook of the environments for scholarly communication and publishing (Section 4.1)?
- What are the different needs and interests of the relevant stakeholders in digital archiving and preservation (Section 4.2)?
- What are the current operational business models for digital archiving and preservation, and how do they compare to those of KB (Section 4.3)?
- What does digital preservation and archiving cost (Section 4.4)?
- Which developments will have an important impact on the future of digital archiving and preservation (Section 4.5)?

Dealing with uncertainty:

- What are the uncertainties and disruptive trends in the future environment of digital archiving and preservation (Chapter 5)?

Relevance for KB:

- What are the short or medium-term and long-term strategic implications of these findings for KB (Chapter 6)?

The inputs for this study to address these questions were based on a series of interviews and a review of the selected literature. More than 50 stakeholders and experts participated in semi-structured interviews either by telephone or in person. (The individuals interviewed for this study are listed in Appendix B) The interviewees were asked to comment on their institutions' interests and needs and their personal views on the current and future situation in relation to the three main themes identified above: scholarly communication, publishing and preservation.

1.3 Structure of this report

The three main themes are reflected in the structure of this report. Chapter 2 provides a background to the concept of digital preservation: the historical context of digital preservation and the e-Depot, the ways in which digital preservation differs from preservation in the print environment and the definitions of several key concepts. Chapter 3 locates KB within the wider landscape of scholarly dissemination and publishing through an assessment of market data, stakeholder positions and emerging trends and uncertainties. Chapter 4 discusses the digital preservation theme: which institutions and systems are active in digital preservation, how the key initiatives compare to one another in various aspects and the relevant developments and uncertainties in this area. Chapter 5 considers the future trends identified in the previous three chapters and proposes four example scenarios for preservation practice, assessing their likely impact on KB. Chapter 6 concludes the report, synthesising the research findings to identify KB's strengths and weaknesses as a player in the digital preservation of scholarly records, and proposes different strategic options to optimise the effectiveness of its activities and alliances.

CHAPTER 2 A brief history of preserving the digital past

This chapter provides a background to the concept of digital preservation. It explains how scholarly communication and publishing have been transformed as a consequence of the digital revolution. This has consequences for the way in which such information must be preserved for future generations. Thus, this chapter presents the concept of digital archiving and preservation in a historical context and explains the role of KB and the e-Depot within that context. Finally, it attempts to define and clarify several key concepts in digital preservation.

2.1 A new era for academic publishing

Publishing in the STM area is an important part of worldwide research output and STM publishers have been at the forefront of developments in the digital era. Research dissemination is dominated by peer-reviewed journals. This is particularly important in STM fields, which receive a relatively large proportion of research funding (see Figure 1). Since the funders of research generally measure the outputs of research by counting the number of peer-reviewed publications and the citations they receive, publishing represents a crucial component of the research cycle.

During the early years of the Internet revolution, several large publishers of academic journals pioneered digital publication and distribution of their journal portfolio. The opportunities presented by the digital age have encouraged publishers to adopt digital delivery and to provide online access to their journals. The majority of journals are now available online, either in parallel with a print version or as e-only (i.e. 'born digital'). Currently the majority of international journals are available electronically, a figure that is reported to be even higher for STM journals, particularly those in English. An important driver of change was Portable Document Format (PDF), which has become the preferred standard for read-only electronic publishing. Adobe's decision to make its Acrobat PDF reader freely available facilitated access to e-journals.

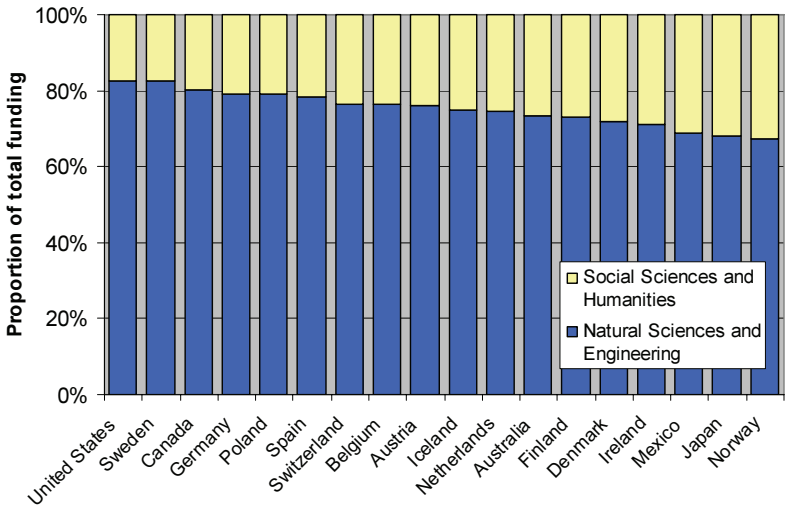

Figure 1. Higher education expenditure on research and development by field of science in selected countries

Source: OECD Main Science and Technology Indicators (2007)

Electronic publishing had created a new business model for the publishing industry, labelled as the 'Big Deal' (Frazier 2001). During the early 1990s, many publishers were confronted with many cancellations of licenses to less popular journal titles. Faced with budget cuts, universities could not bear the continuous inflation of subscription fees. Since electronic publishing is associated with significant economies of scale – the additional costs of delivering an extra copy of an electronic journal are negligible – large publishers introduced package deals to research libraries. Instead of buying licenses for individual journal titles, libraries signed deals for an entire title portfolio, while accepting an increment of approximately 10 percent in the license fee. As a consequence, small libraries which could afford only a small number of selected titles prior to the Big Deal, were the big winners from it. Under these contracts, which are negotiated often by bodies representing consortia of universities, annual price increases were capped for a number of years.

The switch to online publishing has required huge investments from the market players. In order to achieve economies of scale, small publishers almost always have to partner with a third party to facilitate e-publishing. Consequently, the market for STM journal publishing is dominated by several large players, and has been characterised by a trend to consolidation through mergers over the past years (Kobrak and Luey 2002).

2.2 Concerns about perpetual access

With the Big Deal and the trend toward phasing out paper subscriptions, a new problem of digital archiving arose. Library expenditure on printed serials is declining as a proportion of total journal expenditure. Currently, libraries spend approximately similar proportions of their serials budget on print only, e-only and joint electronic and print

subscriptions.[3] While libraries traditionally archived academic output, they are gradually phasing out the print version of their journals. Libraries are unable to afford both print and online formats: they want to save stack space by cancelling print and are also moving into the electronic environment (Waller *et al* 2006). They no longer store these objects in hundreds of square metres of print journal archives, but access them online via publishers' secure servers (such as Elsevier's ScienceDirect). Libraries believe that they own this e-content as a result of licensing conditions and they argue for access to this material in perpetuity. This is a concern for many libraries, since there are a number of possible scenarios, or trigger events – for example, the cancellation or termination of licenses – in which 'perpetual access' would need to be guaranteed (see Box 1).

Waller et al (2006) distinguished nine possible trigger events which could prompt the requirement for perpetual access.

1. When a library has a subscription to a journal that ceases publication.
2. When a subscription to a journal is cancelled by a library.
3. When a subscription to a journal is sold or transferred to another publisher.
4. When the publisher of a journal goes out of business.
5. When a bundle of journals has a fluid title list.
6. When the publisher of a bundle of journals goes out of business.
7. When the publisher of a bundle of journals is bought completely or partially by another publisher.
8. When a subscription to a bundle of journals is not renewed by a library or consortium.
9. When a subscription to a bundle of journals is cancelled by a library or consortium.

Box 1. Trigger events prompting the need for perpetual access

There are various ways to address the requirement for perpetual access. Most licenses have a clause for perpetual access, although there is no guarantee that this will be through online access. Publishers often provide some form of digital media (e.g. optical disk, magnetic tape, CD-ROMs) with back issues, but libraries do not necessarily have the arrangements or procedures in place to facilitate such long-term access. Consequently, libraries are searching for alternative ways to archive digital content to which they have licensed access, but not ownership.

2.3 Digital archiving and preservation

National libraries, like research libraries, are concerned increasingly with the future of digital content. Historically, most countries have a (legal) national deposit function to archive printed publications, often managed by the national library. These collections are preserved in a protected environment that reduces the risk of decay. As the publishing world is moving gradually from print to online publishing, the deposit function of a

[3] In the UK, approximately 37 percent print only, 26 percent e-only and 37 percent joint electronic and print (Electronic Publishing Services, 2006).

national library is shifting as well. As a consequence, national deposit libraries are now building electronic repositories to complement and gradually take over the long-term preservation of, and access to, national print publications. However, the concepts, principles and practices accepted and understood in the print environment may have new meanings or no longer be appropriate in a networked environment (Muir 2001).

First, academic literature in (networked) electronic form no longer has a natural link with its geographic origins. Research is increasingly becoming an international activity through transnational research consortia or international tendering of R&D funds. Furthermore, electronic communication and the international activities that it empowers transcend national boundaries; for example, the shipping costs of digital content have become almost negligible. As international publishers become able to deliver their digital publications from anywhere, the role of archiving in a national context is less clear (Beagrie 2003).

Second, digital media have a relatively short physical lifetime and they become unusably obsolete much sooner. This was recognised in the 1990s, for example, when 5.25 inch floppy disks became obsolete on personal computers (PCs). While the Phaistos disc is said to date back about four millennia and paper can last several hundreds of years, digital media have a much shorter lifespan. Rothenberg (1995) assessed the life expectancy of several media (see Table 1). The contents of most digital media disintegrate long before words written on high-quality paper. Furthermore, digital content possibly faces an even more challenging problem than that of decrypting the purpose and meaning of the Phaistos disc,[4] because of the limited longevity of information carriers and the software and hardware that make the stored information accessible to users. The electronic standards, software and hardware of today will become technologically obsolete in the future.

Third, this leads to a difference with the print environment: not only is the life expectancy of digital media finite, the lifespan of digital formats is also limited. For example, Stanescu (2005) cites the well-known Wotsit's database of formats, which lists some 1,000 digital formats and interim versions, many of which have not been used in more than a decade. Some formats, for example PDF and Tagged Image File Format (TIFF), are standing the test of time. But it is only a matter of time before these formats become obsolete on standard computing platforms. Furthermore, many digital formats must be actively interpreted by software that is runs on a computer in order for their content to be rendered into a form that humans can perceive and use. Such "inherently digital" objects cannot be preserved as page-images or other static representations but must be kept executable indefinitely, even though the original computers and software environments on which their rendering programs originally ran may become obsolete and unusable in just a few years.

[4] See, for example, Balistier (2000), Duhoux (1977) and Trauth (1990).

Table 1. Life expectancy of digital media and the expected time until obsolete

Medium	Average practical physical lifetime	Time until obsolete
Optical (CD)	5–59 years	5 years
Digital tape	2–30 years	5 years
Magnetic disk	5–10 years	5 years

Source: Rothenberg (1995)

Fourth, the trend towards electronic publishing has happened at the same time as the development of increasingly complex additional functionalities that are enabled by a connected and dynamic environment. These functionalities are fundamental to digital preservation, because moving images, hyperlinks with cited papers or access to underlying research data do not necessarily have printed equivalents. An oft-cited example of the consequence of such increased complexity is that approximately 28 percent of the Uniform Resource Locators (URLs) cited in articles in *Computer* and *Communications of the ACM* between 1995 and 1999 were no longer accessible in 2000, and the figure increased to 41 percent in 2002 (Spinellis 2002).

As the national deposit library of the Netherlands, KB was among the first to recognise the issues around digital preservation. As early as 1999 it investigated the possibility of installing a deposit system for digital publications produced in The Netherlands. However, it was recognised soon that for a large category of digital publications, 'imprint' was no longer a meaningful concept. So, the traditional model of a national deposit with geographical boundaries was no longer suitable for guaranteeing the long-term safety of international academic output (van Drimmelen 2004). The following section explains the history of KB's involvement in digital preservation and the e-Depot.

2.4 KB and the e-Depot

As the national library of The Netherlands, KB was founded more than 200 years ago. The aim of a deposit library is to collect published information, preserve it and provide permanent access to the information for use in research, education or for any other purpose in society. In most countries, publications have to be deposited by law, but The Netherlands has a voluntary deposit system based upon agreements between the national library and publishers (Steenbakkers 2003). This has resulted in nearly complete coverage of the print publications produced by commercial publishers in The Netherlands.

With the increasing proportion of born-digital publications, in 1994 KB extended the scope of deposit to include electronic publications and installed a small content management system from ATT (Bell Laboratories). In 1996 KB signed agreements with two of the world's largest STM publishers, Elsevier Science and Kluwer Academic Publishers, on the archiving and long-term preservation of their digital publications with 'The Netherlands' imprint. The pilot deposit system used to facilitate the archiving, based on an IBM product, 'Digital Library', became operational in 1998.[5] As KB soon became aware that, for international journals, an imprint indicating their geographical origin had

[5] For a more detailed overview of e-Depot's history, see Koninklijke Bibliotheek (2002).

lost its meaning in digital form, the Elsevier and Kluwer agreements were upgraded to include all of their journals. The agreements also included back issues (Koninklijke Bibliotheek 2007), since most large publishers had begun to digitise all of their journals back to the first issue.[6]

The first deposit system was a small-scale pilot system, and in 1999 KB issued a call for tender which led to a European tender procedure for the development of a large-scale deposit system for electronic publications. IBM's proposal was chosen from a field of four; following negotiations, the Depot for The Netherlands' Electronic Publications project to build the system began in late 2000. KB identified two distinct main characteristics of this project: development and implementation of a large digital archive, and long-term preservation of digital objects. The latter became the ongoing subject of further R&D rather than a finalised component of the system. After two years of development, on 12 December 2002 the e-Depot, with the IBM Digital Information and Archiving System (DIAS) at its technical heart, was delivered (Koninklijke Bibliotheek 2007).

The functional design of the deposit system was based on the results of the Networked European Deposit Library (NEDLIB) project. NEDLIB, led by KB, was funded by the European Commission and consisted of a consortium of European libraries, publishers and systems developers. NEDLIB contributed to an international standard, under development at that time, for digital archives called the Open Archival Information System (OAIS) reference model (van der Werff 2000). An important design criterion for KB was that the integrity and authenticity of digital publications be kept, i.e. that their original intellectual content be retained and remain accessible. Following implementation of the e-Depot at KB, IBM has collaborated recently with the Deutsche Nationalbibliothek (the German national library) to implement a digital archive labelled 'Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen' (Kopal), based on the DIAS technology.

With the e-Depot now operational, KB has signed archiving contracts with various other publishers. While Elsevier and Kluwer Academic Publishers are international publishers of Dutch origin, the other international publishers represented in the e-Depot are not. Among them are other large publishers, for example: BioMed Central, Blackwell, Oxford University Press, Springer, Sage Publications and Taylor & Francis. KB's target is to include in the e-Depot at least the journals from the 20 to 25 largest publishing companies which produce almost 90 percent of the world's electronic STM literature (Kenney *et al* 2006).

Recent years have seen the emergence of a myriad of initiatives around digitising, storing, archiving and preserving digital media. In an extensive study, the Council of Library and Information Resources (CLIR) in Washington, DC lists no fewer than 12 projects that qualify as legitimate and serious e-archiving efforts. Most of them are based on different technical, financial and institutional models. This is partly because off-the-shelf digital

[6] The first issue of *The Lancet* published on 5 October 1823 became available via ScienceDirect on 8 October 2003. See Elsevier (2003).

preservation archiving systems are not yet available.[7] However, even if such systems were available on the market, the implementations would have to be tailored to the different historic, cultural and geopolitical contexts and the principles upon which each organisation's strategy is based. For KB, these principles can be summarised as follows.

1. *Archiving and preservation of digital objects.* As scholarly output is moving to e-only, a digital archive is needed for KB to continue to fulfil its deposit task as a national library. Archiving and preservation of digital objects is fundamentally different from archiving and preservation of print objects. Archiving and preserving electronic publications for the long term to safeguard future access to their original intellectual content requires a substantial investment in infrastructure, equipment, skills and expertise.
2. *International deposit function.* Since the concept of imprint is no longer valid for digital publications, KB extends its national deposit function to the international level. In so doing, KB offers to become part of a safe places network consisting of a number of digital repositories for international electronic publications. Because of the required scale of investment in equipment, skills and expertise as well as a consequence of publishers' archiving policies, it is expected that there will be a limited number of safe places.
3. *Perpetual access.* KB acknowledges research libraries' concern about the threat of permanent loss of electronic journals and disrupted access to journals for a protracted period following a trigger event (see Section 2.2). An e-Depot would provide a way to manage the threat (Oltmans and Lemmen 2006). Following a trigger event, e-Depot would provide, either temporarily or permanently, affected libraries with access to a specific set of serials and volumes in its archive.

KB's vision of becoming one of the main nodes in the safe places network is untested. It is not yet clear how appropriate the above design principles and their underlying assumptions are.[8] The future of archiving and preservation is highly uncertain and off-the-shelf solutions are not yet available. Because developments in publishing, information services and technology are rapid and dynamic, developing an e-Depot strategy is akin to jumping on to a moving vehicle. However, before assessing this strategy within the wider context of archiving, preservation, scholarly dissemination and publishing, it is useful to understand what exactly these concepts mean.

2.5 Defining some terms

Archiving and preservation are the two terms used most frequently in this report, sometimes separately, often in combination. While archiving and preservation are relatively distinct concepts in the print world, they are not easy to isolate in a digital

[7] Several systems developers are building such systems with a launching customer, however these solutions have yet to be turned into generic, turnkey products. Although these initiatives are distinct, each of them take into account the OAIS standard.

[8] For a summary of the main underlying principles of KB's international e-Depot strategy, see Appendix A.

environment. These terms are discussed below in more detail, and we elaborate on their characteristics and attributes.

2.5.1 Archiving, preservation and repositories

In this report we use both the terms 'digital archiving' and 'digital preservation' when referring to KB's initiative. The term 'archive' has a specific and well-defined meaning in the archival community: in particular, government archives are concerned with retaining formal records of the policies, actions and business processes of organisations and agencies. However, the term is used far more loosely in the library and general information technology (IT) domains. The colloquial meaning of an archive, which we adopt here, is simply a reliable repository of some kind, i.e. a place where documents and other informational artefacts are stored, ideally in a reliable manner that retains their authenticity over long periods of time. Because our interviewees often used the terms 'repository' and 'archives' interchangeably, generally we do not distinguish between them in this report.

Just as the noun 'archives' is used loosely often in the library and IT domains, typically the verb 'to archive' is used to mean little more than simply placing something in a repository. Although archives are often associated with preservation, many so-called archives and repositories have essentially no preservation component other than passive storage. This is compounded by the equally misleading (and widespread) use of the noun 'archive' in IT, where it can mean almost any file or storage structure which contains a collection of information that is no longer actively accessed from its original representation; this again carries no implication of any specific preservation process. Even the OAIS archival reference model says very little about preservation, its 'Preservation Planning' feature having been an afterthought added largely at KB's urging.

While recognising the nearly vacuous meaning of the term 'archive' in the community, we use the term in this report to mean a repository which does concern itself with the preservation of its contents, and attempts to ensure the authenticity of this content for scholarly purposes over timescales of at least decades and ideally centuries. Therefore, we will use the term 'digital archiving and preservation' often in combination when referring to initiatives that provide both access of some kind and preserve authenticity of its content. These two main characteristics are reflected in the definition used by the Research Libraries Group (2002):

> Digital preservation is defined as the managed activities necessary: 1) For the long term maintenance of a byte stream (including metadata) sufficient to reproduce a suitable facsimile of the original document and 2) For the continued accessibility of the document contents through time and changing technology. (p. 3)

However, there is no blueprint approach to digital preservation, as the requirements and specifications depend on the remit and objectives of the parties involved.

2.5.2 Authenticity and integrity

Implicit in the concept of preservation is the notion that what is preserved must remain authentic. As argued in Rothenberg and Bikson (1999), the term 'authenticity' encompasses two concepts: integrity and authentication. The integrity of a preserved object implies that it has not been changed or corrupted in any meaningful way, whereas

authentication is the verification that an object was generated by its purported author or organisation at the purported time and under the purported conditions.[9]

The scholarly record must be authentic if it is to provide a sound basis for scholarly progress, historical accuracy and intellectual and legal accountability. Of the two aspects of authenticity, authentication may be the easier to ensure in the digital domain, since it can rely on the traditional basis of an unbroken chain of stewardship, optionally augmented by digital signature techniques. Integrity may be harder to ensure in the digital domain, since any change to a document may corrupt its integrity.

In traditional scholarly publication, integrity is ensured by preserving originally published copies of documents – or at least accurate page image renditions of those documents which retain virtually all the visual aspects of the originals. Even so, subtle visual and other physical aspects of originals may change or decay over time and may be lost when they are copied as page images. For example, the original colour of inks and paper may fade, ultimately leading to loss of content: witness the US Declaration of Independence, whose original has faded so badly that many of the signatories' names are no longer visible. Additional physical aspects of original documents, such as the chemical and radiological attributes of their paper and ink are lost when they are copied as page images (e.g. photographed). The loss of such features of a document is not often significant, but it can be, particularly when seeking to verify the authenticity and integrity of the document.

In the digital domain, the concept of an 'original' is harder to define than it is for traditional documents. Any copy of a digital artefact that is an exact bitwise duplicate of the original bitstream of that artefact retains all possible relevant digital attributes of the original. Generally, the medium on which the artefact's bitstream was originally stored is not considered a relevant attribute, since a given bitstream may be stored on any number of different media. Only the bitstream itself is relevant for most purposes. Any preservation technique (e.g. migration) that changes the original bitstream of an artefact in any way poses a threat to its integrity, especially if the original bitstream itself is discarded in the process. However, maintaining the original bitstream of an artefact is not enough to ensure that its integrity is preserved; the correct interpretation of that bitstream is equally crucial. Such rendering requires the execution of appropriate software on an appropriate computer. If the original bitstream is not interpreted and rendered in the intended manner, the integrity of the original will be violated.

2.5.3 'Light', 'dim' and 'dark' archives

As mentioned above, storage is ultimately meaningless without access. If no access is ever required, then storage can be avoided entirely. Nevertheless, archives may restrict access to their materials for various reasons, including national security, personal privacy, proprietary or intellectual property concerns, legal or regulatory restrictions, etc. In some cases, such restrictions may have a precise time-limit (e.g. until copyright expires), whereas in other cases they may last until a trigger event occurs (e.g. a company goes out of business or an individual dies). Access restrictions may be:

- absolute – no one is allowed to access the material;

[9] For a more in-depth discussion of authenticity in the context of digital preservation, see Rothenberg (2000b).

- partial – some content may be edited; or
- role-based – subscribers to a publication series or members of a library may be granted selective access to archived material.

Repositories that prevent all access (at least for some period of time) are referred to as 'dark' archives. However, in most cases some access is required, if only to perform administrative functions such as cataloguing, restructuring or verifying the integrity of the content of a repository. In such cases, limited user-based or role-based administrative access may be granted; in other cases, selected users may be granted access to selected materials in an archive under restricted conditions (for example, on library premises). In such cases a repository may be referred to as a 'dim' archive. Conversely, repositories that provide full access are referred to as 'light' or 'open' archives.

CHAPTER 3 Scholarly dissemination and publishing: a complex and dynamic environment

This chapter locates KB within the wider landscape of scholarly dissemination and publishing through an assessment of market developments and figures, stakeholder positions and emerging trends and uncertainties.

3.1 The outlook of scholarly dissemination and publishing: key figures and trends

The importance of scientific publishing lies in its role in the selection, production and spread of scientific and technical knowledge (European Commission 2007). Not only is this an important public function, enabling dissemination of knowledge to foster economic growth and further research, but it is also a thriving commercial market.

3.1.1 STM publishing market

The size of the global STM publishing market (which includes journals, books and secondary information services) is estimated at around €7 billion.[10] A small number of large market players dominate the STM publishing market. In 2003, 66 percent of global revenues were generated by the eight largest publishers (Elsevier, Thomson, Wolters Kluwer, Springer, John Wiley, Taylor & Francis, Blackwell and American Chemical Society) (see Figure 2). The profit margin for an established journal is estimated around 35 percent. With the substantial economies of scale yielded in the digital age, it is likely that the trend toward consolidation will continue for some years. For example, Springer Science+Business Media has been the fourth player in the market ever since private equity investors bought BertelsmannSpringer and Kluwer Academic Publishers and merged them in 2003 (Simba Information 2007). Wolters Kluwer cited lack of scale as the reason for opting to exit scientific publishing and focus solely on medical publishing (Gooden *et al* 2002).

[10] However, estimates differ considerably. According to Simba Information, global science and technology revenue was $7.6 billion, up 4.5 percent from $7.3 billion in 2005.

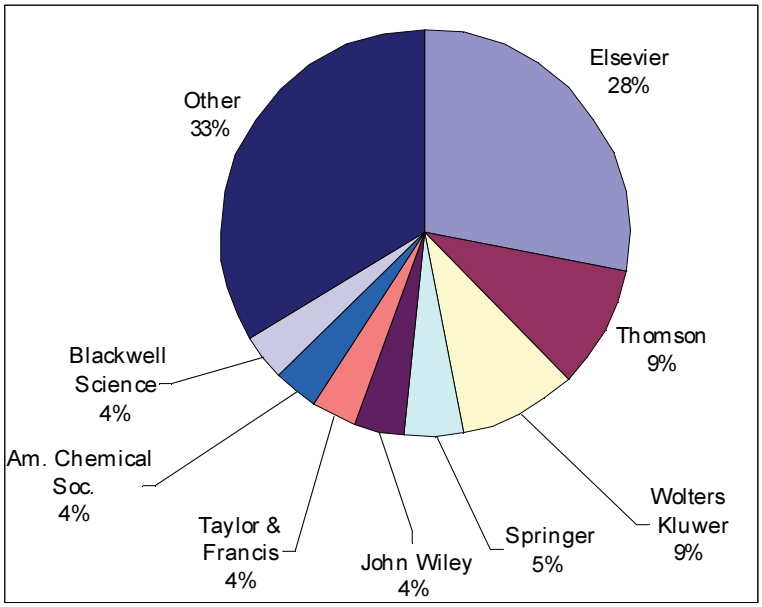

Figure 2. Global market shares in STM publishing, 2003

Source: Powell (2004)

The total European STM publishing market is estimated at between 24 and 32 percent of the total STM publishing market, namely €1.7 to 2.2 billion in 2004 (Electronic Publishing Services 2006). According to the OECD *Science, Technology and Industry Outlook* (2006), R&D expenditure on public and private R&D in OECD countries was €586 billion, while the European Union (EU) 25 spends approximately €200 billion (European Commission 2007). The European Commission estimates that, combined with the figures on the publishing market, journal subscriptions account for some 0.6 percent of the total European R&D budget.

Journal sales account for the lion's share of the revenues; estimates vary between 35 and 50 percent (see Figure 3). Consequently, the market value for journals is estimated at between €2.7 and €3.5 billion. However, the importance of journals differs across disciplines. Journal articles are relatively more important in the sciences and social sciences; books and monographs are more important in the arts and humanities (European Commission 2007). In the humanities, 50 percent of the library budget is spent on periodicals, while in STM 70 percent is spent on periodicals.

With the shift from print to electronic serials, journals have been able to continue dominating the market. Currently, 60 percent of journals worldwide are available electronically and this figure may be as high as 90 percent for English language journals (Electronic Publishing Services, 2006). Opportunities from digital technologies and networks have been taken up at a slower pace in the humanities, in part due to cultural barriers (European Commission 2007). By 2016, it is estimated that half of all serial publications will have migrated to electronic-only format, and that STM titles will be the first to switch (Electronic Publishing Services 2006). Large publishers will start with their

less profitable titles.[11] Smaller publishers, especially learned societies, will switch on the basis of rising print and distribution costs (Powell 2004). In a recent paper, Karen Hunter from Elsevier acknowledged that there are four issues that publishers face before they can move to electronic-only formats; among these issues is 'bullet-proof digital archiving of electronic journals' (Hunter 2007).

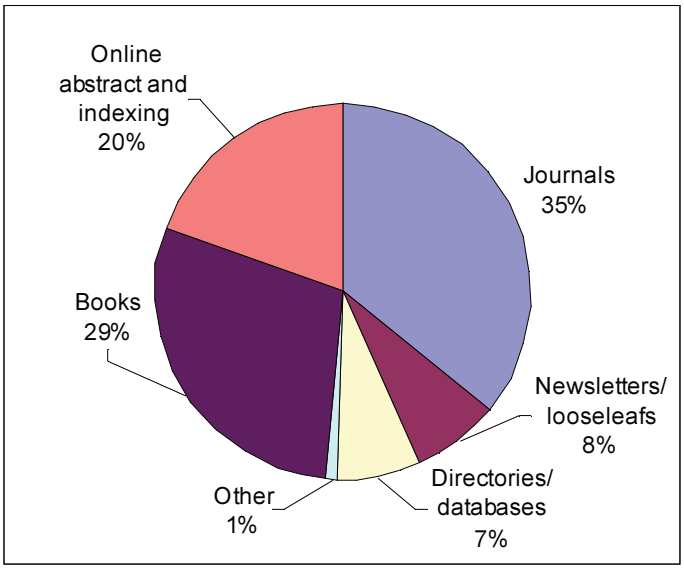

Figure 3. STM publishing market breakdown by delivery format, 2006

Source: Simba Information (2007)

How many journals actually exist worldwide is debatable. The total depends on the inclusion criteria for the research fields, language areas, types of serials and quality assurance. As at 28 June 2007, Ulrich's database states that there are 23,202 active, peer-reviewed scholarly and academic journals. According to Mabe (2003) there are many caveats in estimating this number, but this is more accurate than the 'hundreds of thousands' sometimes quoted by other researchers. With temporary exceptions, for most of the last three centuries, the growth rate of active, peer-reviewed scholarly and scientific journals has been around 3.5 percent per annum. This means that the number of active journals had been doubling every 20 years (Mabe 2003). An extrapolation of this trend would imply that by 2050, there will be more than 100,000 titles (see Figure 4).

The number of articles has grown significantly over recent decades. Currently, the annual output in peer-reviewed journals is estimated at 1.4 million articles (Mark Ware Consulting 2006). Among the large STM publishers, market leader Elsevier currently publishes around 250,000 articles a year (approximately 25 percent of all STM articles published), accumulating to a total of 8 million articles over its long publishing history (interview with Nick Fowler 2007).

The majority of scholarly publishing output comes from Europe and the USA. According to the European Commission (2007), there are 780 publishing houses based in Europe

[11] Wiley-Blackwell's forecast suggested that in the sciences, 39 percent of journals would be online-only by the end of 2007 (JISC 2007).

which are responsible for publishing 49 percent of all research articles; 43 percent of the world's research papers are published by European authors, and it is estimated that Europe accounts for 24 to 32 percent of world expenditure on journals European Commission (2007).

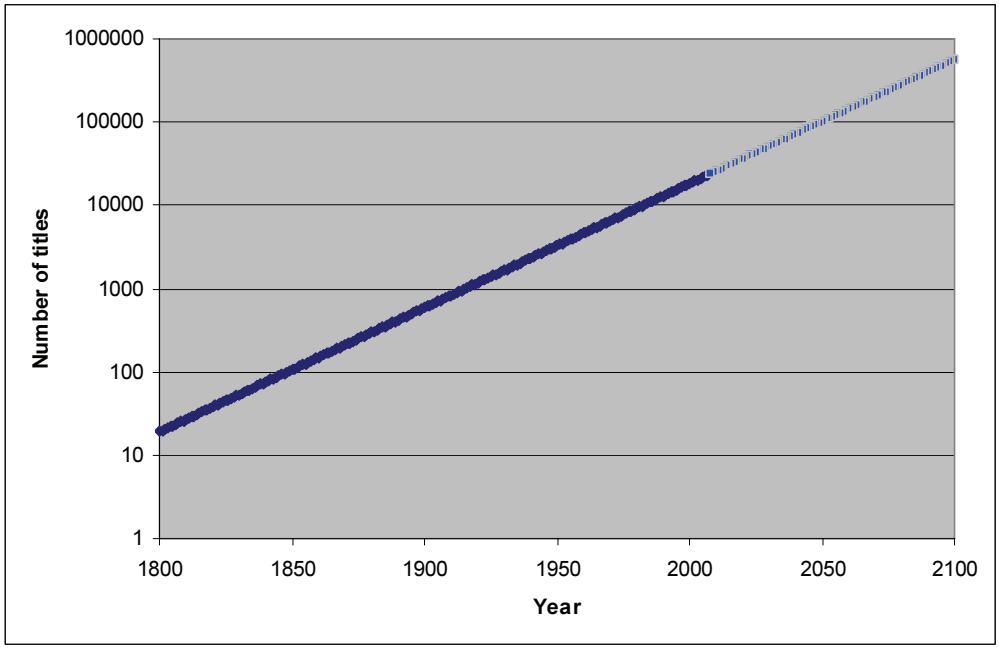

Figure 4. Growth of the number of academic journals over time (historic and projected, based on a 3.5 percent annual growth rate)

3.1.2 Costs and pricing of publishing

The prices of journals and the profits made by publishers have been subject to continuing criticism from the academic and library community. Prices are perceived to be high and the annual license cost increases have not received a warm welcome. In the USA, for example, the Association of Research Libraries reported that the annualised price rise for journals during the period 1984–2004 was 7.6 percent, compared with the US Consumer Prices Index annualised rise of 3.1 percent over the same period (Mark Ware Consulting, 2006). While subscription costs continued to increase, library budgets remained relatively constant. Such price inflation has led to subscription and license cancellations. Such declines in circulation have led in turn to further price increases (Tenopir and King 2000).

There are considerable differences between different fields. For example, the average annual price for a journal in physical chemistry is €2,189, while a title in sociology costs €325 on average. In addition, a study by Dewatripont *et al* (2007) shows that within fields there are price differences between commercial 'for-profit' publishers and not-for-profit publishers (typically, learned societies and academic publishers). Generally, the price per article (see Figure 5) and journal prices (see Figure 6) are set higher by for-profit publishers.

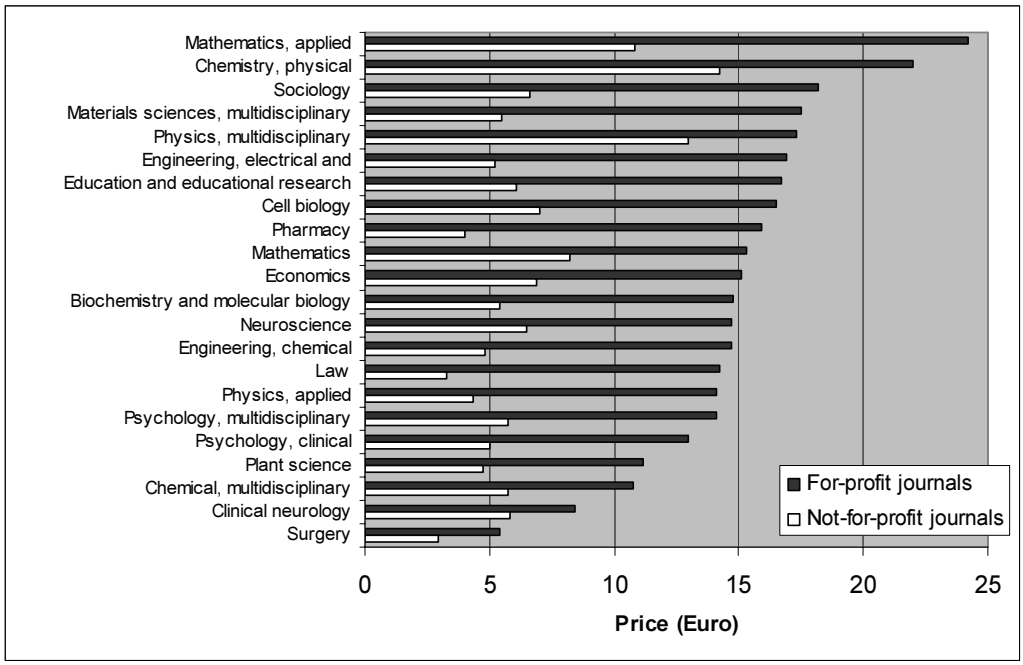

Figure 5. Average article prices from for-profit and not-for-profit publishers

Source: Dewatripont *et al* (2007)

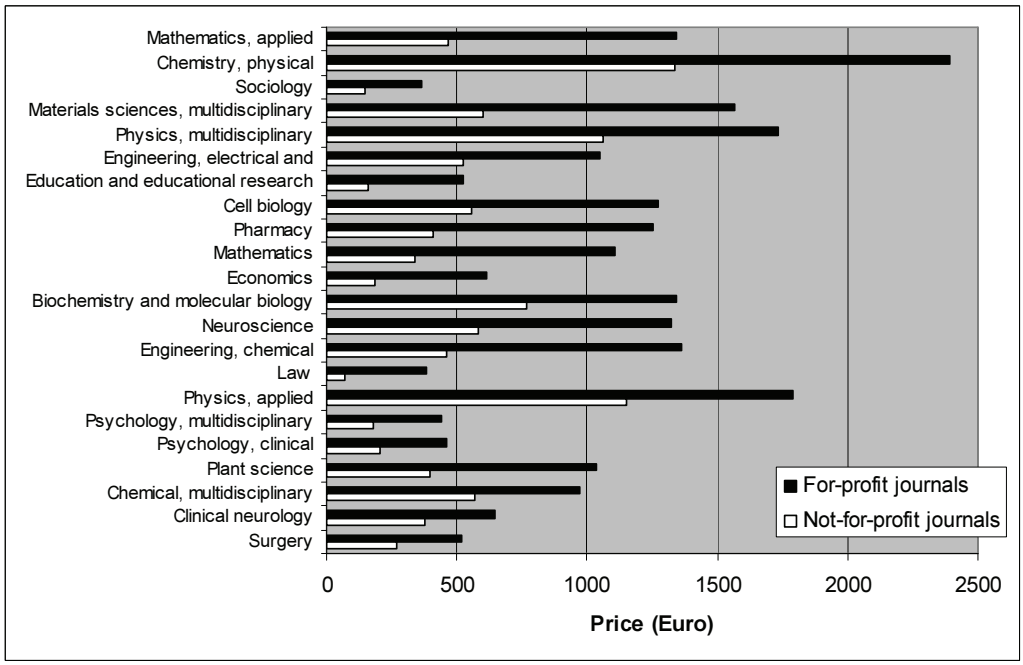

Figure 6. Average journal prices from for-profit and not-for-profit publishers

Source: Dewatripont *et al* (2007)

In open-access models, published content is freely available, but publishers charge their authors for the costs related to editing, peer review, distribution, etc. BioMed Central, for example, charges €1,000 per article for most of its journals and PLoS charges €1,200 (European Commission 2007). Similarly, other publishers are beginning to offer open-access options in traditional journals. However, thus far, the level of uptake of open access for hybrid journals is relatively low: typically in the range of 4 to 8 percent of the total number of papers published (interview with Nick Fowler 2007).

Although open-access publishers generally have to supplement their funding with other sources of income such as advertising and donations, BioMed and PLoS believe that they will break even soon. However, Butler (2006) argues that PLoS lost US$1 million in 2005 and its author fees and advertising revenues covered only 35 percent of its total costs. Nonetheless, supporters of open access believe that within a reasonable time, all pure research papers will be open access.

3.1.3 Lifetime value of journal publications

e-Journals have existed for only about a decade. Furthermore, most publishers have digitised their back files only recently and some have still to embark on this cumbersome task. Therefore, it is difficult to estimate the value of digital publications over a longer period. Usage of print articles may be a proxy for this value. Although changes in availability and access introduced by electronic publishing may change patterns of journal usage, Tenopir and King (2000) reported that the age of articles read by university scientists in a given year (between 1993 and 1998) was remarkably similar to the results that they reported in a study performed in 1960. The result of their 2000 research is presented in Table 2.

Download statistics from Elsevier also suggest that the usage of older publications (see Figure 7) is not dissimilar to that reported by Tenopir and King. Although there are significant differences between subject areas, publications older than 15 years constitute less than 5 percent of the full text downloads in a given year. However, with the improved online accessibility of older content – ranging back to their first volumes – the lifetime usage of scholarly publications may well be subject to change over the next years.

Table 2. Proportion of readings by age of scholarly articles by university scientists 1993–1998

Age of articles access (years)	Proportion of total (%)
1	58.5
2	12.3
3	6.2
4–5	7.7
6–10	9.3
11–15	1.5
15+	4.6

Source: Tenopir and King (2000)

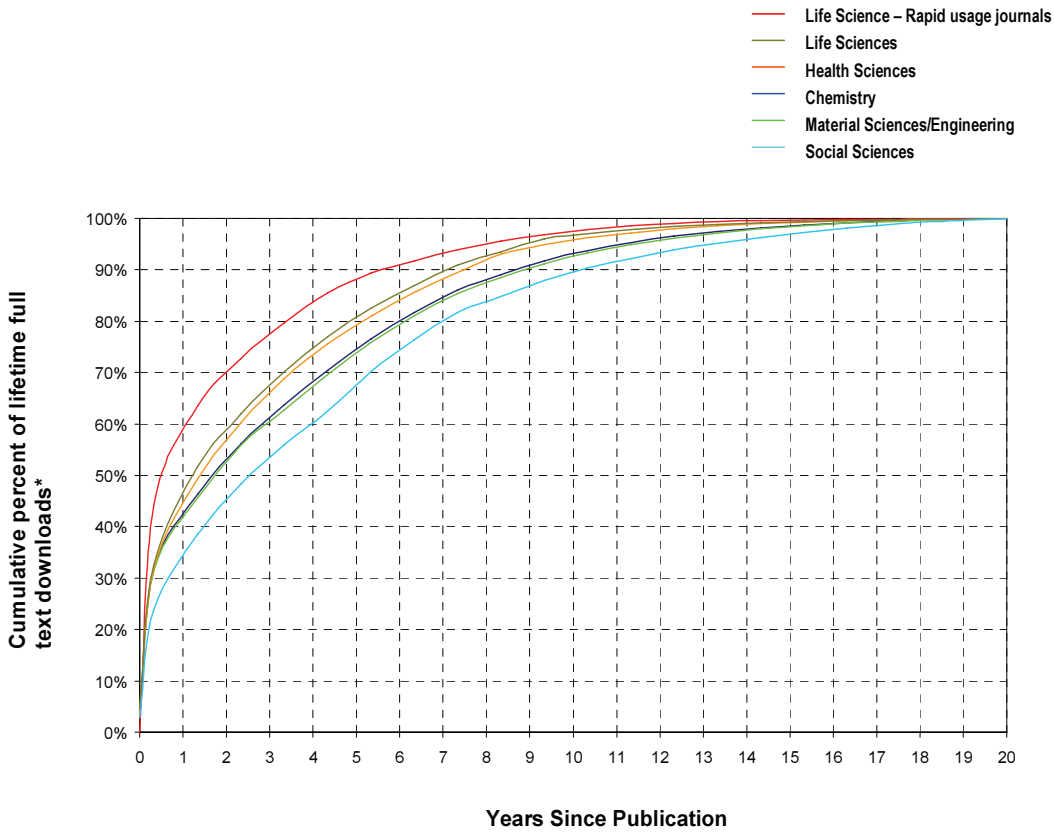

Figure 7. Usage statistics by subject area[12]

Source: Elsevier Science (2007)[13]

Because usage statistics are not widely available, citations are often used as a proxy for the lifetime use of both print and digital publications. Although various caveats have to be taken into account when using citation statistics as a proxy for usage,[14] data from Thomson Scientific's[15] citation index suggest that half the citations in any year refer to material older than six years. This indicator, cited half-life, can be used to compare the value of journals over time of different journals, fields and even publishers. Here we only draw attention to the finding that the half-life of high-impact journals has increased over the last decade; relatively more articles cited relatively old papers (see Figure 8). This points to a possible trend break from the relative constant lifetime usage of journals according to Tenopir and King. This trend can be labelled as the increasing long-tail value of publications.

---

[12] For purpose of analysis, this assumes article lifetime usage is only 20 years; typically, the actual lifetime of article usage is significantly longer than 20 years.

[13] Email communication with Karen Hunter, Elsevier Science, 6 June 2007.

[14] Some scholars have tried to address these caveats. See for example, Moed (2005).

[15] The largest database of journal publications and their citations is owned by Thomson Scientific (formerly Thomson ISI), which calculate the journal impact factors. Journals listed in this database are referred to as 'ISI indexed'.

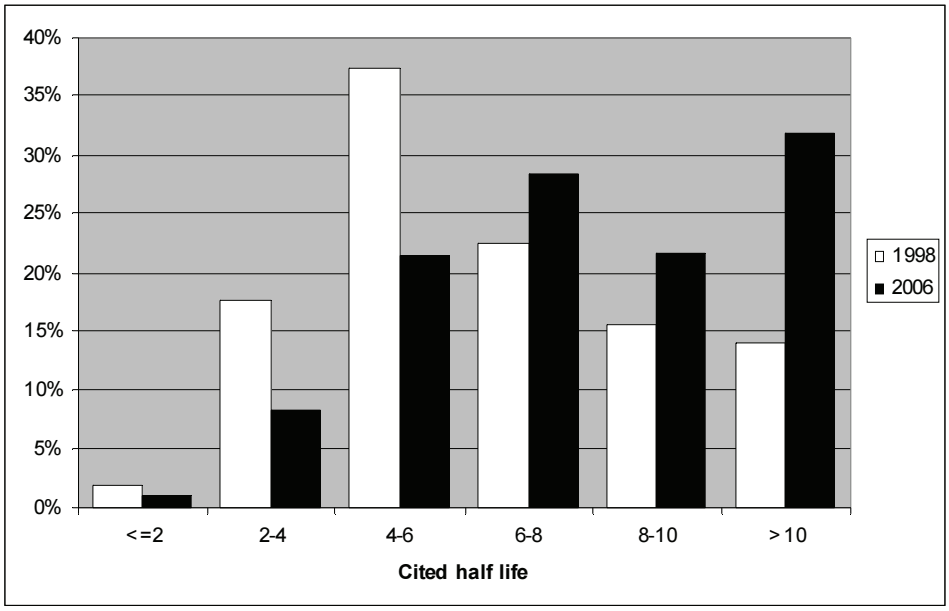

Figure 8. Cited half-life distribution of the top 500 journals in *JCR Science Edition*

Source: Thomson Scientific (2007)

3.2 Stakeholders' views on scholarly dissemination and publishing

In order to understand the issues and developments in scholarly dissemination and publishing, it is useful to understand the different interests, needs and objectives of all the stakeholders involved in these areas. (These positions and interests are discussed in more detail for each stakeholder group in Appendix E.) Based on a series of interviews and available surveys and literature, we briefly touch upon the views and positions of the key stakeholders.

3.2.1 Demand and supply of scholarly output

While researchers and authors are responsible for the supply of intellectual content, they also constitute the demand for academic content, mostly via their libraries. Publishing is the premier way of disseminating the results, but funding and furthering the author's career are underlying drivers. In order to be able to 'stand on the shoulders of giants',[16] researchers need to have access to the work of those giants. Peer review is still the fundamental mechanism underlying the quality of scholarly output; the importance of this mechanism is underlined by the academic community (Rowlands and Nicholas, 2005; Rowlands *et al* 2004). Therefore, in addition to depositing papers in an institutional repository, scholars have a continuing interest to publish research results with traditional publishers. The priority for retaining copyright still seems to be relatively low (Rowlands *et al* 2004).

[16] Isaac Newton's modest reflection on his achievements, famously expressed in a letter to Robert Hooke in February 1676: "If I have seen further it is by standing on the shoulders of giants."

As reported by the House of Commons Science and Technology Committee (2004), libraries are atypical consumers. Rather than purchasing more goods until the benefit they receive is balanced by the cost, they spend up to the limit of their budgets. If prices rise, libraries will purchase fewer journals, if prices fall they will purchase more. Similarly, if research output rises but library budgets remain unchanged, libraries will cancel the least popular journals. The ceiling on budgets means that publishers of 'must-have' journals hold a monopoly position, and can increase their revenues when they raise prices, as lesser journals are discarded by libraries. This explains why observers from university libraries expressed little confidence in the motives of large publishers.

3.2.2 Traditional publishing model

STM publishing is a large market with a limited number of large players and a large number of small players. Traditionally, scientific publishers undertake a series of functions that have emerged for efficiency reasons: peer review; copy editing and typography; database preparation; production and distribution; and archiving (Boyce, 1996). A large professional service provider can execute these tasks much more efficiently than authors.

The traditional business model of journal publishing is based on license fees for title subscriptions, which changed into package licenses after the Big Deal. It is often argued that the market for STM publishing is, in many ways, an imperfect one. Competition between journals is limited, as no two journals are exact substitutes; all journals are monopolies in a certain niche market and research libraries, often represented by umbrella organisations for consortia of universities, have weak negotiating power when faced with increased subscription fees. In contrast, publishers argue that publishing has inherent costs; they add value to the research process by warranting the quality of journal articles (through the editorial boards which oversee the academic quality of journals) and distributing them efficiently (interview with Paul Ayris 2007).

> *'We add considerable value to the research process by organising, appointing and managing journal editorial boards, investing to promote and market journal brands, coordinating the input of over two million manuscripts globally each year, managing the quality control process of peer review, producing journal articles in both print and electronic formats, distributing journal articles globally in both print and electronic formats as efficiently as possible, providing ongoing updates to articles (e.g. comments, corrections, withdrawals) and preserving them in perpetuity.'*
>
> Nick Fowler, Elsevier Science

As a result, publishers rely on university libraries' and professional associations' continuing need for high-quality scholarly papers, and ensure quality mainly by means of professional external peer review. Their business model is built on research institutions' ability to fund and willingness to pay for maintaining library collections.

The sustainability of this traditional model is a subject of intensive debate. Some, predominantly the traditional publishers, argue that the journal article as a definitive object of research dissemination will continue to dominate scholarly communication. This model will continue to work as long as the peer-reviewed published journal article is the accepted, quality-assured object of academic dissemination. Subscription cancellations are likely to focus on peripheral journals with small audiences. JPMorgan believes that companies such

as Elsevier, with strong catalogues of ISI-indexed journals, remain well positioned to outperform the market (cited in Mark Ware Consulting, 2006).

There are diverging opinions on the urgency of assuring perpetual access. Publishers generally claim that there is no access problem, while libraries increasingly demand guarantees for perpetual access that go beyond a promise in the license agreement. Additionally, learned societies are starting to use archiving arrangements with an independent third party as a criterion for selecting a publisher to produce their journal. Following such demand from the market, publishers are entering into agreements with long-term archivers. In fact, most of the members of STM (the International Association of Scientific, Technical and Medical Publishers) have signed the so-called 'Brussels Declaration' (STM 2007), which includes statements on perpetual access and access to research data.

3.2.3 Alternative publishing models

In recent years, the open-access movement has claimed to be a reasonable alternative to the traditional publishing model.[17] In the 'author pays' system, subscriptions are free, but the author is charged for copy-editing, peer review and distribution. This charge is usually paid by a research grant or through institutional funds.

> *'The traditional scientific publishing market is an imperfect market. Not only do a few large publishing companies dominate the market, but if readers need access to a particular research article, they cannot typically make use of an alternative article from another journal as a substitute. … Open-access publishing makes better use of the underlying economics of digital communication by taking advantage of the near-zero cost of dissemination to provide universal access.'*
>
> Matthew Cockerill, BioMed Central

Open access is viewed often by research libraries and funders as more sustainable than traditional publishing, since it can be linked more directly to research funding; dissemination can become a distinct component of a research grant. In addition, the costs incurred in scientific publishing in an online environment are mostly proportional to the number of articles published, not to the number of readers (interview with Matthew Cockerill 2007). The UK research councils and The Wellcome Trust, for example, have set open-access requirements for papers arising from research that they have funded (The Wellcome Trust 2006). Similar shifts are evident in Canada and France. The US National Institutes of Health (NIH) funding of research generates 60,000 articles a year and NIH are expected to shift to an open-access policy as well (interview with Robert Kiley 2007).

Self-archiving is another element of open access, enabling authors to disseminate their research articles for free over the internet and helping to ensure the preservation of those articles in a rapidly evolving electronic environment (Joint Information Systems Committee (JISC), 2005; House of Commons, 2004). Self-archiving or institutional open-access repositories such as PubMed (NIH), UK PubMed (The Wellcome Trust) and university digital repositories could potentially harm publishers' revenues in the longer

[17] For a comprehensive overview of stakeholder views on open-access publishing, see Albert (2006).

term. Thus far, these alternative models have not had a big impact on STM publishing and scholarly communication.

3.2.4 Public deposit function

Various countries have specified a legal deposit function, which obliges publishers to deposit their publications in one or more national repositories. Recently, national libraries have begun to acknowledge that, due to the increasing proportion of electronic publishing in scholarly communication and in order to continue fulfilling their public responsibility, they should invest in digital archiving services. Once the initial investment has been made, these services should be scalable to include a wider range of content and provide value-added services which may generate additional revenues and/or strengthen the national libraries' public good function. However, in Germany, where a preservation system (Kopal) has been designed specifically for meeting the needs of a national deposit function, it indicates that a 'narrow' deposit function model still exists. Keeping in mind that the German context is different from that in The Netherlands,[18] the Deutsche Nationalbibliothek takes a narrower public service task and focuses solely on publications in the German language and those about Germany published abroad in other languages (interview with Ute Schwens 2007).

Other stakeholders' perceptions of national libraries vary considerably. The libraries' long history is viewed often as an important asset, creating confidence that they will continue to exist in perpetuity. Furthermore, their not-for-profit status and public role avoid suspicion of an underlying economic motivation. The public interest is best served by libraries, in order to ensure that all publications are collected and preserved in perpetuity. There is no apparent commercial or other reason why a for-profit organisation would incur cost to guarantee longevity of data and continued accessibility if there is no financial reward for doing so. However, there are criticisms of libraries: the distinction between preservation and access is not always clear. KB is facilitating on-site access to its e-Depot content for its 'walk-in' users and does not require a separate license agreement for such access. Herewith, KB has come to shift towards facilitating research in other areas than its historic remit. Whereas KB traditionally focused on the humanities and social sciences, it has moved slowly into the territory of traditional science and technology libraries (e.g. the Delft University of Technology library). This is perceived by some libraries as an ambiguous situation (interview with Maria Heijne 2007); KB's focus on archiving digital publications is perceived by some universities as a move into becoming a research-facilitating library (interview with Kurt de Belder 2007).

Furthermore, being funded by the country's taxpayers and having a 'national library' stamp may create the perception that a national library will only act in the national interest; this may work against the library's aspirations in the global arena, where national boundaries are becoming less relevant.

[18] For example, Germany and the German-language area have a much larger scale than The Netherlands and the Dutch language. Furthermore, the German federal governance structure consisting of 16 states (*Bundesländer*).

3.3 Trends and uncertainties in scholarly dissemination and publishing

The publishing environment has experienced several drastic developments over the last decade. Entering the digital era and the Big Deal are just two obvious examples. These developments have changed the remit and context of deposit libraries and raised concerns about perpetual access. Changes in the publishing world are causing a redefinition of the issues that will affect digital preservation of scholarly output in the future.

A number of trends appear to be emerging in the ways that scholars produce, disseminate and use information in the digital age. As pointed out by Harley *et al* (2006), scholarly behaviour varies across disciplines, according to disciplinary cultures. However, it seems likely that at least some of these trends will diffuse eventually across discipline boundaries. First, we discuss the trends themselves and then explore their implications for KB.

3.3.1 Increasing use of 'grey' (informal) publication

Many of our interviewees pointed out that informally-distributed information has always played a role in research. Letters, pre-publication drafts, workshop talks, etc. have provided previews of results and enabled authors to obtain feedback on their work before publishing it. However, information and communication technology (ICT) has expanded the role of such information greatly by enabling it to be 'published' via email, blogs, websites and self-publication services (such as Lulu).[19] Coupled with the open-access movement,[20] this has created an explosion of 'grey literature' which is now relied upon in many scientific disciplines (and some humanities disciplines as well) to disseminate results prior to – and sometimes instead of – publishing them formally. This may lead to a future reawakening of scholarly interest in the differences between distinct versions or editions of published works (paleography), as was common in the humanities before the advent of the 'lens' of standard editions in the 19th century.[21]

Also, institutional and discipline-based publication may be increasing. These may produce formal, peer-reviewed publications such as those of traditional publishers, or they may produce grey literature with a lower or different standard of review. In any case, they may use novel ICT-enabled distribution mechanisms in addition, or in preference, to printing. Reinforcing this trend is the growth of institutional repositories within many university libraries; these are being developed to make locally-produced informal (and formal) publication available to scholars within these institutions themselves or to other users of these institutions' research resources. Discipline-based repositories, including open-access repositories such as PubMed Central and arXiv, also appear to be expanding, serving as single-point sources for information within their disciplines. However, as noted in Kenney *et al* (2006), open-access repositories are unlikely to contain complete collections of scholarly literature in the foreseeable future; additionally, despite the names of these repositories (notably arXiv), few of them are concerned with long-term preservation.

---

[19] See: http://www.lulu.com/

[20] As recorded in the Berlin Declaration, available at: http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.

[21] This possibility was suggested by Amy Friedlander of CLIR who, in turn, attributes the core of the idea to Chuck Henry (also of CLIR; interview with Amy Friedlander, 2007).

Further, local repositories are seen as a growing trend; by focusing on materials of local origin or interest, these serve as a way for local institutions (including public and private libraries) to justify their investment in building collections as they become incapable of competing with the large-scale universal collections that are available ubiquitously online. Finally, personal digital repositories are seen as a growing trend; these are analogous to personal library collections, which traditionally have included informal material such as manuscripts, letters from colleagues, etc.[22]

Grey literature is seen by some as a bifurcation in the publication process. Traditionally, formal, peer-reviewed publication has served to support the tenure and promotion process, as well as providing the formal record of progress in science, but some observers see both of these roles as being usurped to some extent by informal publication. As Harley *et al* (2006) point out, although there is still considerable scepticism about the possibility of performing tenure and promotion by alternative means, this is a possibility. Eventually, mechanisms such as citation and usage counts or distributed online peer review may supplement or supplant traditional peer review as the accepted standard.[23] Similarly, grey literature may replace formal publication to some extent as the *de facto* scientific record, especially if the preservation and archiving of informal publications can be ensured. In contrast, observers representing large traditional publishers still have a strong belief in the sustainability of the traditional journal as the main dissemination vehicle for science.

3.3.2 Increasing reliance on information hubs and 'pull' vs 'push'

Scholars are relying increasingly on Internet information hubs, such as Google and Yahoo!, to enable them to search for what they need. This is seen as part of a broader shift from information 'push' (represented by formal publication and journal subscription) toward information 'pull' (users search for what they want). Whereas the older, push paradigm makes scholars the passive recipients of information, the newer, pull paradigm casts them as active information seekers. At the same time, other ICT developments are pushing information to scholars in new ways: person-to-person email, mailing lists, listservs, 'publish/subscribe' mechanisms and 'bots' (robotic software agents which can search for desired information) provide new ways of alerting scholars to the availability of new information which is likely to be relevant to them.

In addition, data mining is gaining popularity in many fields. Although it is analogous to traditional research, its use of automated and semi-automated tools to search huge databases and corpora of literature produces a qualitatively new capability. Data mining may be performed in advance as a (push) service by publishers or other information providers to create useful datasets, or it may be performed on-demand (pull) by researchers

[22] The growth of such institutional, discipline-based, local and personal repositories is an interesting counter-trend to the spread of universal, distributed, virtual repositories and information hubs, which are accessible without geographic restriction. It remains to be seen whether specialised repositories can carve out meaningful niches for themselves or whether they represent a futile attempt by specialised organisations to justify their traditional roles as collection builders.

[23] One observer commented that the traditional peer-review process gives excessive weight to the opinions of a small number of reviewers (typically three), who often have ulterior professional or personal motives for their assessments.

themselves. In either case, one of its salient characteristics is that it accesses disparate information resources which may span traditional authorship, publication and ownership boundaries. This implies a need for virtual corpora consisting of federated, linked information that resides in what traditionally have been separate institutions, disciplines and publication venues.

3.3.3 Increasing expectation of online access

As scholars increase their reliance on digital resources, they expect relevant information to be available online. Whereas it may have been sufficient a few years ago to provide online indexes to offline resources, the expectation now is that the resources themselves will be accessible online. Among other things, this makes scholars less aware of and concerned with which institutions own or offer what information. Online access creates a virtual merging of disparate information sources. Scholars will always be concerned with the authenticity of the information that they obtain, which may depend on the reliability and credibility of its stewardship, so institutional credentials may continue to be important from this perspective. But which institution(s) own, link, combine, package, present and provide access to information is likely to be seen as increasingly irrelevant in the face of rising virtualisation.

Furthermore, this virtualisation may impact one of the most deeply-held traditions among libraries: namely, their emphasis on the development of their collections. On the one hand, in a fully-linked and virtualised online environment, the collection of an individual institution may become irrelevant to its users, except to the extent that this collection can be transparently merged with complementary collections of other institutions to provide a virtual universal collection. On the other hand, searching a universal virtual collection of this kind may be quite inefficient, since scholars may find it harder to restrict the results of their searches to the relevant subsets. This problem is related to the need for metadata to create a 'semantic web' (World Wide Web Consortium 2001), since free-text search in an unlimited context tends to produce many spurious results. Therefore, institutional collections and focus may continue to serve as a proxy for semantic metadata in the short (and possibly medium) term.

3.3.4 Increasing reliance on underlying data and models

Especially in STM, there is a definite trend toward the increasing use of data and models underlying published results. Scientists are disseminating and using each other's datasets, models, lab notes, etc. to help analyse, reproduce and review each other's work (the ability to reproduce being essential to the scientific method). This is seen by many as a significant paradigm shift to a new, deeper level of sharing and analysis, which has the potential to accelerate and improve the quality of scientific research greatly. As is often the case, the social sciences and even the humanities are expected to follow this trend in the coming years, at least to some extent.

Moreover, in many cases (such as the Human Genome Project and space-based earth observation projects), the primary results of a scientific project are data rather than analytic reports. In such cases, although published reports may summarise the results, the valued output of the project lies in the data that it produces. Such products may take various

forms, but they are most likely to consist of databases, complete with pre-packaged queries and executable programs that provide appropriate and useful access to the data.

The forms of these traditionally unpublished materials are often quite different from those of published reports, or even informal drafts and preprints. Often, datasets are represented in databases rather than in printed tables, whereas models are embodied in executable, interactive computer programs. Lab notes may be in any of a number of forms and may include multimedia documentation, such as photographs, video, or instrumentation traces. This is seen by some observers as greatly complicating preservation and archiving, since the digital forms of these materials are unlikely to be simple page-image formats such as PDF. In many cases, the relevant formats are inherently digital; that is, they are executable (i.e. programs) or heavily encoded and so must be run as (or rendered by) software in order to be used. Furthermore, the informal nature of these materials may not allow them to be forced into the relatively small set of standard formats that typically are required for formal publication (although the need to be able to share this informal information itself may provide some pressure to standardise their formats).

3.3.5 Increasing dissemination of novel types of material

As noted above, the desire to disseminate and use underlying data, models, lab notes and other supporting documentation for scientific projects as well as the data output of such projects leads to an expansion of the range of formats and types of digital materials that must be handled. Reports and articles are peppered increasingly with links to other reports, articles, datasets, models, etc. that reside on the web. At the same time, individual units of information may contain multiple, compound, embedded subunits, such as charts, graphics, images, animations, video clips and other multimedia components. Some universities are encouraging graduate students to produce multimedia e-dissertations, which may embed executable components such as simulations or active or interactive visualisations.

Many informal documents also link to dynamic components such as databases or geographic information systems, which are being updated continuously, or to active web pages or other dynamically-generated content which does not exist in static page-image form. In addition, active links (such as URLs) enable scholars and researchers to access finer-grained units of information such as individual sections, chapters, paragraphs or tables within larger works. Ultimately, this may lead to an information space in which larger units (such as reports) exist only as virtual collections and combinations of smaller, 'atomic' units (for which our current paradigm does not even offer an accepted name). Furthermore, each component of a compound document (whether the whole exists physically or only virtually) may be represented in a different digital format, requiring a distinct rendering program. Again, this expanded range of formats complicates preservation and archiving.

Finally, although as yet these new types of materials have not challenged the traditional book significantly, e-books are gaining gradual popularity and eventually may become a new form in their own right. e-books are already popular forms of scholarly dissemination in the humanities and large publishers have indicated that the business model for e-journals is being adapted gradually to series of e-books. Particularly in STM e-book series with chapters by different authors, book chapters are slowly gaining a status that is similar

to that of journal articles (interview with Peter Hendriks 2007). Even if e-books and chapters are thought of as self-contained digital objects which are not dynamically linked to other objects on the network, but are instead carried and viewed as distinct entities, they may still embody most of the features above, including internal links, multiple components in different digital formats, underlying self-contained databases and executable programs. This expanded version of the current e-book form would be nearly as difficult to preserve and archive as a web-based document.[24]

3.3.6 Increasing production of executables

A significant trend that is implied by the increasing reliance on underlying data and models and the increasing dissemination of novel types of material is the use of executable programs in scholarly communication. Many of the new types of informal material that are being disseminated and used in the scientific communities (and even in the humanities) use executable or interpretive formats which can be rendered only by running a program on a suitable computer.

The dissemination and use of models and simulations constitutes a further step in the sophistication of scholarly communication in the sciences. Often, scientific results are based on the use of models or simulations, so making these available in executable form enables other researchers to verify and reproduce results, or to run 'excursions' to try other cases and explore the range of applicability of a claimed result. Models and simulations can be arbitrary interactive programs written in any programming language and intended to run on any computer. Therefore, preserving such programs in executable form (so that they can be used by future researchers) presents serious challenges.

Finally, many fields in the sciences and humanities are using animations, videos and interactive computer-generated visualisation to display their results in forms that are understood more easily and have greater impact than is possible on the printed page. Rotatable 3D models of molecules undergoing chemical reactions, fly-through models of virtual ancient cities or extra-terrestrial landscapes, animations of time-based trend analysis and similar techniques are becoming gradually widespread in the hard sciences, social sciences, history, archaeology and other fields. Products such as these constitute a form of informal scholarly communication that relies on the ability to disseminate computer programs in executable form.

Preserving such inherently digital products is particularly challenging, since they may depend heavily on the visual and other perceptual characteristics of display screens and other peripheral devices that are used to interact with them and render their output.

3.3.7 Increasing reliance on new methods of review and evaluation

Traditional peer-reviewed formal publication remains the backbone of the tenure and promotion system at most universities and the foundation of professional advancement in most scholarly fields. However, some observers have begun to consider the possibility that this situation may change, as the bifurcation of the publication process widens. As informal publication increasingly becomes the basis for substantive interaction among researchers

[24] KB has indicated that it recently signed an agreement with Elsevier for archiving its e-books series, and Springer has indicated its interest to sign a similar agreement.

and scholars in many fields, formal publication may be relegated to the role of supporting professional advancement.[25]

If this bifurcation proceeds as current indications suggest, the increasing marginalisation of formal publication may lead to the emergence of informal mechanisms for performing peer review and supporting tenure and promotion and professional advancement. (Some observers note that the university tenure system itself is under attack from many quarters, having already given way in many places to the expanded use of non-tenured faculty and may not survive much longer, making tenure and promotion a non-issue.) Alternative peer-review mechanisms based on informal publication are easy to envision and may have some advantages over traditional peer review (for example, as noted above, the current system of peer review gives inordinate weight to the opinions of a very small number of reviewers who often have vested interests that influence their evaluations). Informal peer-review mechanisms might be based on reviews submitted by solicited or self-selected qualified readers, or on citation indexes or usage patterns, which can be automatically recorded in the digital environment. Mechanisms such as citation counts are already serving as informal evaluation mechanisms in many fields, even where formal peer review is still in place. Although most established (e.g. tenured) researchers and scholars remain reluctant to consider the possibility of replacing traditional, formal peer-review mechanisms with new informal ones, some are already entertaining such possibilities (Harley *et al* 2006).

3.3.8 Increasing long-tail value

A final trend which may have significant impact on the publishing industry, and thereby indirectly on libraries and the scholarly communication process, is an increasing interest in the long-term revenue potential of published artefacts. ICT facilitates the repackaging, reuse, mining and repurposing of previously published material. Realising this, scholarly publishers are beginning to think of their published materials as assets with long-tail value – a point of view which already pervades the music and film industries. This may tend to make publishers far more hesitant to entrust their assets to libraries or others without restricting access. Such control may include the creation of 'dark archives' (Pearce-Moses 2005), which become accessible only after some trigger event (see Section 2.2).[26]

The extreme extension of copyright protection reinforces this trend by ensuring that publishers will be able to continue to profit from their assets in the future. Also, universities and other institutions may see the potential for longer-term returns on their publication investments, which may fuel the observed increase in institutional publishing and the development of institutional repositories. Finally, authors may recognise this long-tail value potential, which may lead them to be more cautious about signing away the rights to their works in order to publish them formally, especially since self-publication is becoming an attractive option.

---

[25] The other primary role of formal publication is the provision of the formal record of progress in each field, but even this may be taken over by informal publication, especially if informal products can be preserved for the long term.

[26] However, not all publications possess equally long-tail potential. For example, textbooks are often revised and republished every few years, negating the value of their previous versions.

In essence, open-access publication is incompatible with this increase in potential long-term value, because once something is released as open source, presumably it cannot serve as a revenue generator. Furthermore, it may be difficult to prevent others from repackaging and reselling open source material, reaping its long-term benefit at the expense of the author. Although informal publication is not necessarily open access, it too is at odds, at least in its present form, with reaping long-term value. However, alternative 'pay-per-view' or usage-fee models may enable informal publication to generate revenue, as is happening in stock photography.[27]

3.3.9 Implications for KB

Many of the trends and developments discussed above are in their infancy. Some of them may never become significant, whereas others may not become so for years or decades. For example, scholarly communication still relies heavily on formal publication, but institutional repositories are becoming more important for disseminating research results.[28] At this moment, most formal publication still consists of static page-images, such as journal articles and books; the vast majority of the works that are being published currently, both formally and informally, rely on static text as their main constituent, making preservation relatively easy. Nevertheless, some or all of these trends seem likely to produce far more complex digital scholarly products in the future. In particular, objects that are inherently digital will include, or rely on, executable code which must be run on a computer in order to render the object viewable or usable by a human. The long-term preservation and archiving of such objects in usable form is likely to be highly problematic.

> *'The roles of authors, publishing and libraries will shift, so the way we go about preserving material will need to adapt to these shifts as well.'*
>
> Amy Friedlander, CLIR

KB appears to be one of the very few institutions in the world today that is taking a truly long-term view of preservation, including the preservation of inherently digital objects. This may make its mission somewhat difficult to justify to other institutions that are focused on short-term approaches, but it is recognised by many as having thought more deeply about long-term issues than most other institutions. In order to retain and deepen this long-term perspective, KB will need to continue to monitor the above trends and developments and assess their implications for its current and future plans and directions. There are a number of implications that appear to be relevant at this moment. However, given the rapid rates of change in the ICT, publication, library and preservation domains, these and other potential implications should be re-examined and re-evaluated on a frequent basis.

[27] However, granular pricing of this kind may be more difficult to predict – and therefore budget – than traditional subscriptions. Furthermore, one observer noted that since traditionally, scholarly publishers have made most of their money by selling aggregations, granular pricing would involve a significant paradigm shift for them.

[28] KB indicates that it is already active in this area through formal relations with the institutional repositories of the Dutch universities for archiving and preserving published material.

CHAPTER 4 Digital archiving and preservation: an area under construction

The successful long-term archiving and preservation of digital scholarly communication depends on a number of factors. These include:

- establishing appropriate institutional mandates, agreements and funding models;
- forging appropriate relationships among publishers, libraries, archives, universities, information hubs and others;
- appropriate public support, appropriately recognising and adapting to commercial interests and concerns, developing and deploying appropriate preservation metadata; and
- developing and adopting suitable technical mechanisms to ensure that digital objects will be authentically and usably accessible in the distant future.

Although all of these factors affect the outcome of the preservation enterprise, having a sound technological basis for digital preservation is a very fundamental one. Appendix C elaborates the technological basis for preservation.

The following section discusses the current outlook of the preservation environment. It will focus on organisational and institutional issues of preservation, assess the main existing models and describe trends and uncertainties for the future.

4.1 The outlook of digital preservation: key figures

Over the past centuries, institutions such as libraries and archives have taken the collective role of preserving national and collective heritage. Beagrie (2002: 10) labelled them the custodians of the "collective memory". Often, other institutions' involvement in this field has been for much shorter time horizons. Beagrie argues that it is not surprising that memory institutions have been the first to identify the challenges associated with digital preservation. However, as the core mission of knowledge institutes and libraries in particular has come to depend on access to digital objects, archiving and preservation has become an issue beyond that of cultural heritage.

Table 3. Summary of 12 included digital archiving initiatives

Name		Start year	Type of entity	Principal access mechanism	Designated users
CISTI Csi	(CSI)	2000	National science library	Current online (partners)	Partners
CLOCKSS[29] (Controlled LOCKSS)	(CL)	2006	Limited membership/subscription	Trigger or audit	Available to everyone after a trigger event
KB e-Depot[30]	(KB)	1996	National library	Trigger or audit (plus on-site), may provide online access (open) to open-access content	Everyone in The Netherlands (since 1996) and beyond (since 2002)
Kopal/DDB	(KOP)	2004	National library	Trigger or audit (plus on-site), moving wall planned	Patrons of the Deutsche National Bibliothek and the Goettingen State and University Library
LANL-RL	(LANL)	1995	Government/ limited subscription	Current online, plus on-site to the general public	LANL staff and subscriber staff
LOCKSS (Lots of Copies Keep Stuff Safe) Alliance	(LA)	2005	Open membership	Trigger or audit	Local users served by library maintaining LOCKSS box
NLA PANDORA	(NLA)	1996	National library	Current online (open)	Australians and anyone with a research interest in Australia
Online Computer Library Center (OCLC) ECO	(ECO)	1997	Open membership/ subscription	Current online	Subscribers to the ECO collections
OhioLINK EJC	(EJC)	1998	Limited consortium	Current online	Supporting members of the Ohio higher education community
Ontario Scholars Portal	(OSP)	2001	Consortium	Current online (members) based on what they have licensed or purchased	Institutional members of the Ontario Council of University Libraries
Portico	(PORT)	2005	Open membership/ subscription	Trigger or audit	Members, at least initially
PubMed Central	(PMC)	2000	National medical library	Current online (open) and moving wall	Users of the National Library of Medicine in the USA and beyond

Source: Kenney *et al* (2006)

Probably the most up-to-date overview of the current landscape of preservation initiatives of digital scholarly communication is provided by the 2006 CLIR report 'E-Journal Archiving Metes and Bounds: A Survey of the Landscape' (Kenney *et al* 2006). This study focused on "the 'who, what, when, where, why and how' of significant archiving programs operated by not-for-profit organisations in the domain of peer-reviewed journal literature published in digital form" (Kenney *et al* 2006: 10). To avoid repetition of this comprehensive survey, we briefly summarise the main findings of this study

As digital archiving is still a new field, most initiatives have focused on the selection, acquisition, storage and maintenance of digital collections. Only now are the efforts

[29] This entry has been adjusted for reasons of accuracy: CLOCKSS is initially a dark archive, but after a trigger event, its content is to be made freely available to anyone.

[30] This entry has been adjusted for reasons of accuracy.

needed for long-term preservation beginning to be identified and addressed and it is not immediately clear what a digital preservation initiative is, and what it is not. Kenney *et al.* did not include digitisation efforts (e.g. JSTOR, library-led digital conversion projects or the Google digitisation project, self-archiving efforts by publishers and initiatives still being planned). The 12 archiving initiatives surveyed by Kenney *et al.* included arrangements for perpetual access and e-journal archives, distinguishing between maintaining access rights beyond subscription periods and mitigating the risk of permanent loss to ensure availability for future users (see Table 3). The most successful initiatives noted in the survey were located at institutions which had been working on practical implementations and policy for several years.

Kenney *et al.* summarise the archiving programs from the perspective of North American university libraries. They conclude that programs have secured their mandates, defined access conditions and are making good progress toward obtaining necessary rights and organisational viability. However, room for improvement is particularly apparent in the areas of content coverage, meeting minimal services and establishing a network of interdependency. Furthermore, few (if any) of the surveyed initiatives are addressing long-term preservation. Appendix D includes some relevant results from this survey in comparative tables.

4.2 Stakeholders' views on digital preservation

Opinions on the need and objectives for digital archiving and preservation are not necessarily similar. Views are largely dependent on the stakeholder's perspective, as publishers' interests differ from those of university libraries. Even within stakeholder groups opinions vary depending on, for example, nationality, field differences, etc. This section discusses these views and summarises stakeholders' needs for, and interests in, preservation. Appendix F provides a more elaborate overview of these perspectives.

> *'There is a lack of awareness for the strategic value of designated infrastructure solutions for digital preservation.'*
>
> Ute Schwens, Deutsche Nationalbibliothek

With a few exceptions, most institutions and scholarly communities appear to be ignoring long-term preservation issues, focusing instead on short-term solutions that may last no longer than five or at most 10 years (interview with David Prosser 2007). This appears to emanate from a belief that the long-term problems are intractable, so worrying about them now will prevent making any progress. However, some argue that if these problems are not addressed, any short-term approaches that are adopted are likely to fall far short of future needs.

> *'We are not really concerned about preservation at the moment, beyond to the market demands. … If a new standard will be adopted in future, this is a new opportunity for us to sell our content again.'*
>
> Andy Williams, Cambridge University Press

However, as discussed in Section 3.2, research librarians are concerned increasingly with perpetual access. Therefore, when negotiating new license agreements, clauses on arrangements with third-party digital archives become an important aspect of negotiation. It is not entirely clear what an ideal third party for digital preservation would be. Most observers agree that there should be a number (between five and 15) of trusted digital archives that are geographically distributed, allowing for overlap and different technological strategies. This largely fits the safe places model.

The services provided by these third parties would function as an insurance policy for perpetual access. Despite the lack of a coherent vision, some in the scholarly community claim that there are certain requirements for an organisation to be credible as an insurance provider. First, several observers commented that, in order for any organisation to be a credible candidate for performing preservation, preservation must be part of its core mission. Because traditionally, publishers have not considered preservation to be part of their business models, they are considered by many to be unlikely to do a good job of preserving scholarly material. The general perception is that publishers may be coerced into preserving material, for example, to guarantee future access to the e-journals that they publish, but that preservation holds little positive value for them, only the avoidance of negative value (i.e. avoiding litigation by subscribers who can no longer access back issues of an e-journal). Even the possibility of new long-tail revenue is not seen as sufficient to make preservation part of the core missions of most publishers: it is thought that the uncertain potential for such future revenue would be too weak an incentive.

An organisation's motives must be trustworthy for it to serve a credible preservation role. For-profit organisations such as publishers or information hubs may be seen as having many potential conflicts of interest stemming from their commercial goals. (However, most insurance companies are profit-making organisations.) Publicly-funded and non-profit organisations are seen to be more reliable in this regard, although non-profit organisations may not be trusted implicitly either, depending on their funding source.

> *'Does it have to be a not-for-profit solution? If a for-profit organisation has a sustainable, viable, trusted and cost-effective business model, it should be considered as an option.'*
>
> Steven Hall, Wiley-Blackwell

Surprisingly, similar views were expressed by several research librarians. Trustworthy insurance provides a contract with explicit policy conditions. One librarian noted that if libraries had to pay for this service, it would improve their confidence that the insurer would fulfil its moral (or even legal) obligations to provide access in case of a trigger event. In contrast, some librarians argue that universities should not be responsible for the costs of these digital archives. Observers commented that if journal license agreements contain explicit phrases on perpetual access – specifying that publishers are responsible for providing access for universities in case of a trigger event – archiving costs should be included already in the license fee (interviews with Matthew Cockerill and Kurt de Belder 2007).

Furthermore, some observers remain sceptical about not-for-profit initiatives such as Portico, due to its partial financial dependence on publishers for its continued existence.

National libraries appear to be among the candidates trusted for preservation, although even among these, some (including KB) seem to be more respected than others. However, some observers expressed concerns that eventually, there may be a conflict between the national interests of national libraries and the needs of the international community (interview with David Prosser 2007). According to this reasoning, an international consortium could be an optimal solution.[31]

Yet another aspect of credibility is an organisation's credentials as a steward for scholarly communication. This involves the ability to provide long-term assurance that the authenticity of the scholarly record will be maintained. Authenticity has a technological aspect (as noted above), but most observers seemed more concerned with the dangers of intentional human intervention, whether resulting from monetary, intellectual, ideological or political motives. Here again, national libraries seem to be among the most trusted of such institutions, since they are perceived as having the fewest potential conflicts of interest. However several observers noted that even governments have not always been above attempting to modify the scholarly record, so the best arrangement would be a consortium of national libraries which maintain replications of each other's repositories and conduct periodic cross-audits of each other's holdings.

Finally, several observers expressed concern that key preservation functions should not be outsourced. Even if preservation is a core mission for an organisation, it may decide to outsource some aspects of the process. This is seen as a questionable strategy, since it may reduce the organisation's intellectual investment in preservation or reduce its competence in performing the process. For a library, outsourcing preservation is seen as tantamount to outsourcing the library's collection, which is seen as its core asset and therefore something which should not be entrusted to others.

This scepticism about the motives, credibilities and competencies of various organisations seems to overshadow technological issues for most observers. Organisational and business model concerns seem paramount to many in the community, with technology a distant second. However, this may be due in part to the fact that few observers are thinking about long-term preservation. Instead, they are focusing on very short timeframes of no longer than about five years, during which time technological issues probably can be ignored, especially with regard to page-image artefacts such as current e-journals.

4.3 Assessment of three preservation models

In this section we will discuss the current business models for the preservation of scholarly communication. We describe each organisation, strategy and funding model, describe the

[31] The Research Libraries Group – National Archives and Records Administration (RLG–NARA) taskforce on digital repository certification recently conducted a study on Trustworthy Repositories, published by the Center for Research Libraries and Online Computer Library Center (Research Libraries Group, 2007b). A test audit of the KB e-Depot, using the RLG–NARA audit checklist for the certification of trustworthy digital repositories and other metrics, was performed by RLG–NARA on 25–26 April 2006 and the results were presented to KB recently in 'The KB Final Report, Draft March 21, 2007'.

advantages and disadvantages of each design and technology, assess their strengths and weaknesses and summarise stakeholders' views of these systems.

4.3.1 Three organisational approaches to preservation

As discussed above, Kenney *et al* (2006) analysed 12 e-journal archiving programmes from the perspective of concerns expressed by directors of academic libraries in North America. They divide these programmes into three groups:

- government-sponsored – primarily funded by national libraries (KB's e-Depot, the British Library's Digital Object Management (DOM) programme and the Deutsche Nationalbibliothek's Kopal;
- consortia – which aggregate content primarily for access but have assumed archiving responsibility (OhioLINK EJC and the Ontario Scholars Portal); and
- member or subscriber initiatives – most of which have been launched specifically to address digital archiving issues (CLOCKSS, LOCKSS Alliance, OCLC ECO and Portico).

Although the e-journal focus of the Kenney study emphasises issues involving traditional publishers, many of its observations and conclusions are relevant to the broader subject of digital preservation as a whole.

The categorisation of these three groups of approaches corresponds broadly to sources of funding, not to specific business models or distinct preservation strategies. This is no fault of the study, but rather a reflection of the variation and uncertainty that is present in the missions, business models, access policies and preservation strategies of the many emerging so-called 'digital archiving' programmes. In fact, the study draws no significant conclusions that are related to these groupings. However, on the one hand, the authors do note that "Programs with a government mandate may have an edge in terms of ongoing commitment and funding appropriations, although an exclusive dependence on government largesse could be detrimental in lean economic times" (p. 61), and that those programmes whose primary missions include the provision of access "may also be at a financial advantage, because the costs of archiving are tied directly to current use and subscriptions" (p. 62). As a corollary, they suggest that programmes which are not government-funded and are intended primarily for preservation (e.g. CLOCKSS, LOCKSS Alliance and Portico) may be the most vulnerable.[32] On the other hand, current indications are that institutions are hedging their bets by entering into agreements with multiple programmes: although this may be a good strategy in the face of uncertainty, it makes it difficult to tell which programmes have the broadest base of support.

The following three sections present an assessment of current approaches to the preservation of digital scholarly communication which seem worth analysing and comparing. All three of these are in their infancy and so cannot be evaluated yet in terms of their efficacy, viability or reliability; nevertheless, they are widely recognised as the leading efforts in the field so far. The findings of this assessment build on the available

[32] However, as noted below, CLOCKSS or LOCKSS are not focused on preservation per se, rather on ensuring perpetual access.

documentation on these models and existing meta-evaluations without duplicating them. Therefore, in these sections we have focused on the stakeholders' views on these models.

4.3.2 e-Depot

KB's e-Depot became operational in 1996 and its technical infrastructure was scaled up to the current size in 2002, becoming one of the first digital archives focused on long-term preservation. In subsequent years, its coverage of international scholarly publications has grown substantially, to more than six million digital objects in March 2006, representing 3,500 e-journal titles, in about 6 terabytes of storage space (Kenney *et al* 2006). According to Table 4, which provides a detailed overview of the objects ingested in e-Depot, the content of e-Depot will exceed 10 million digital objects by the end of 2007.

Table 4. Objects ingested in e-Depot as of 1 August 2007

Publishers	2003	2004	2005	2006	2007	Total
Elsevier	1,518,904	967,985	1,690,317	2,766,510	973,663	**7,917,379**
Kluwer Academic Publishers	0	295,524	39,474	27,939	201,187	**564,124**
Springer	0	0	6,664	59,572	379,313	**445,549**
NTvG (Dutch publisher)	0	0	291,302	2,189	1,002	**294,493**
BioMed	0	0	96	18,731	4,178	**23,005**
International Union of Crystallographers (IUCr)	0	0	0	5,471	71,075	**76,546**
Brill	0	0	0	23,886	1,746	**25,632**
Oxford University Press	0	0	0	356,515	53	**356,568**
Blackwell	0	0	0	0	0	**0**
Taylor & Francis	0	0	0	0	0	**0**
Sage	0	0	0	0	0	**0**
IOS Press	0	0	0	0	0	**0**
CD-ROMs	350	355	3	336	71	**1,115**
Members of the Dutch Publishers Association[a]	0	0	0	81	151	**232**
Digital Academic Repositories (DARE)	0	0	0	66898	26233	**93,131**
e-books	0	0	0	0	0	**0**
Web archive	0	0	0	0	0	**0**
Total	**1,519,254**	**1,263,864**	**2,027,856**	**3,328,128**	**1,658,672**	**9,797,774**

[a] The objects included under 'Members of the Dutch Publishers Association' stem from a small number of Dutch publishers depositing under the agreement with the Dutch Publishers Association. They do not include objects from Elsevier, Kluwer Academic Publishers or NTvG.

Source: KB (2007)[33]

KB's e-Depot initiative is of great interest to many observers, due to its situation within a respected national library which has had a recent history of progressive thinking about digital preservation. Stakeholders note that KB, as the national library of The Netherlands, has a quality mark associated with it. In contrast to national libraries in France, Germany and the UK, KB is seen by most observers as neutral, internationally oriented, not ideologically biased and associated with a country that is known for its willingness to work

[33] Personal communication with Hans Jansen, Koninklijke Bibliotheek, 26 August 2007.

with others. Nevertheless, some observers feel that KB's national affiliation makes it less credible as an international preservation agency. KB's government funding is seen primarily as an asset, although it also may be a vulnerability (in times of national retrenchment). Additionally, KB's relation with a national government is seen as a potential limitation, since scholarly communication knows no national boundaries. Furthermore, as informal digital publication becomes more important, linked, distributed virtual objects will make national borders – and geography in general – increasingly irrelevant. Therefore, many observers question whether a small government's library, funded by taxpayers, will be able to justify the potentially unbounded task of preserving the scholarly record, particularly in the international arena.

Some observers feel that libraries in the USA will always be US-centric and demand a US solution to perpetual access; in which case, KB may not appeal to them. However, those observers interviewed in the USA know about KB and respect both its reputation and its efforts in preservation.

> *'I doubt that KB will be accepted by American libraries as the guarantor of their perpetual access. It is likely that US libraries will want a US solution to the issues of preservation and perpetual access. As further solutions develop KB may play a largely domestic archiving role. In the longer term it is likely that publishers like Blackwell would prefer to deal with a small number of archives than with many.'*
>
> Steven Hall, Wiley-Blackwell

KB's model involves voluntary agreements with scholarly publishers, which enables the library to provide free access to non-open-access deposited works as well on library premises. Such a 'dim archive' arrangement allows a certain level of ongoing access and, as a consequence, there is some random testing of the viability of the archived information. This limited onsite access is part of all the archiving arrangements with the publishers archiving at KB. In the light of recent extensions of copyright protection as well as the possible increase in publishers' perceptions of long-tail value from the repackaging, reuse and repurposing of their published works, it is not clear whether such onsite access will continue to be available without further restriction, and this is all the more problematic for the online access that is increasingly demanded by scholars and researchers.

Therefore, KB's access model is questioned by some scholars, who worry that agreements with publishers may limit future access (whether or not publishers contribute financial support to KB's efforts). At the same time, the model is questioned by publishers who worry that KB's mandate and its orientation as a library towards scholars and researchers may motivate it to provide greater access than publishers may want. However, in the case of open-access publications, publishers urge KB to make the publications as widely available as possible. Additionally, if in the hypothetical case that KB is forced to compromise access by preserving published works in a 'dark archive' (this is not KB's policy or its intention), its role as a national library may be called into question. As a result, some observers feel that KB must explore new access and rights management policies. Furthermore, some observers (notably publishers) doubt that KB's promise to open its content to the world after a trigger event is operationally feasible, or even necessarily desirable. Finally, some observers question whether preservation without continued access (as in a 'dark archive') can be reliable, since it would not continually test and verify the

accessibility, usability and authenticity of preserved objects. For example, although The Wellcome Trust's PubMed does not have explicit long-term preservation techniques, being openly accessible it may have a greater chance of surviving, because it is viewed by hundreds of thousands of people every day.

Uncertainty about KB's funding model leads some observers to question the long-term viability of its preservation model. It is unclear who would pay for KB's efforts in response to a trigger event. Currently, KB cannot guarantee that its servers would sustain temporary open access to a very large number of e-journals in the case of a trigger event. For example, access to Elsevier's content would exceed KB's capacity. One solution to this would be to sign an agreement with an IT service to provide extra server capacity immediately after the occurrence of a trigger event. In KB, arrangements with the archiving publishers are included that in such an event, the publisher agrees to compensate KB for the direct costs. Most interviewees agree that it is unlikely (and probably infeasible) that the service provided by KB would continue to be free. In particular, libraries would not expect access after a trigger event to be free: research libraries already buy many services from commercial providers and this is seen as just another such service.

> *'While in the development phase it would have been less prudent to charge for its services; now that it has established e-Depot and a leading preservation facility, KB should do what it can to develop other funding mechanisms.'*
>
> Cornelis van Bochove, Netherlands Ministry of Education, Culture and Science

In general, the cost of hosting e-Depot is substantial and the service it provides is at least in some ways similar to a commercial back-up service for publishers. It may ultimately be unreasonable to expect the Dutch taxpayer to take responsibility for supporting such a service, especially if its scope is to be international. In order to reduce risk (and thereby increase its credibility), KB may need to diversify the funding of e-Depot beyond its traditional government support base. However, it may be difficult for KB to justify charging publishers for preservation, especially since this service has been provided free of charge up to now; conversely, charging scholars for access would seem to violate KB's role as a national library. One observer noted that insurance policies are trusted more if participants pay a fee; similarly, without some kind of contractual agreement, a preservation agent such as KB cannot be held accountable for loss, which limits its credibility.

> *'What is an insurance policy without a premium?'*
>
> A librarian

KB's unique, multi-pronged technical strategy for preservation includes migration, the use of IBM's Universal Virtual Computer (UVC) to perform 'data preservation' and the use of emulation. Only the first of these (migration) is familiar to most people, so the latter two are viewed with some suspicion and confusion. Many observers remain highly sceptical about emulation and assume that it would be more expensive than migration.[34] In

[34] This assessment appears to be based on little empirical evidence; the empirical evidence that exists, in fact, suggests the opposite. For more information on the costs of preservation and the differences between emulation

summary, KB's technical approach to preservation is considered as among the most advanced in the world; however, its reputation as an international service provider is far less established.

4.3.3 Portico

Begun officially in 2005, Portico is a non-profit membership organisation aimed at the robust preservation of e-journals. Its funding model relies on membership fees from publishers and libraries, in addition to some initial funding from JSTOR, Ithaka, the Library of Congress and the Andrew W. Mellon Foundation. Although it is relatively new, it already has (at the time of writing) one of the most complete sets of agreements with publishers and committed e-journal titles: 38 publishers from the USA and elsewhere representing a commitment of about 6,000 journals and more than 360 libraries from the USA and eight other countries.[35] This robust set of agreements makes it very attractive to libraries. Portico's business model consists of holding published e-journals in a 'dark archive' and allowing library members to access these holdings in the future if a trigger event occurs. (This archive may be 'dim' rather than 'dark' in the sense that some access may be permitted in some cases.) Portico accepts e-journal source files from publishers, rather than capturing published images such as PDF files representing distributed e-journal articles. It claims that it is five times as expensive as e-Depot, although its fees are not excessive for most publishers. Libraries contribute between US$1,500 and US$24,000 per annum, while publishers' fees range from US$250 to US$75,000 per annum, based on their total journal revenues (Fenton 2006). Although an annual fee of $75,000 is substantial, it is not an excessive amount for a multi-billion dollar business. However, a contribution of US$24,000 for a library may be excessive for many libraries. Portico is heavily oriented toward the USA, which may appeal to US universities, but it has done an excellent job of international marketing. Most of the libraries interviewed which have an agreement with Portico are satisfied. Traditionally, libraries have had good experiences with and confidence in JSTOR and support the activities of the Andrew Mellon Foundation; consequently, Portico is a logical trusted party.

All of the content that Portico receives is "normalized" or pre-emptively migrated prior to being preserved in the Portico archive. Two migrations are performed on all submission packages: the first to re-package the publishers' content into a METS-based archival unit,[36] and the second to migrate the bibliographic and full-text content.[37] These two migrations allow Portico to reduce diversity as it actively manages the ongoing preservation of the

---

and migration, see Section 4.3. and Appendix C. In addition, the ongoing Dioscuri effort, a joint project by the KB and the Dutch National Archives, is developing an emulator specifically designed for preservation purposes, which may help convince sceptics of the viability of the approach.

[35] As of the September 11, 2007, 2.09 million articles or 31.7 million digital objects that total approximately 2TB have been ingested into the Portico archive (personal communication with Eileen Fenton 2007).

[36] METS is a standard for encoding descriptive, administrative, and structural metadata about objects within a digital library, expressed using XML. METS is being developed by the Digital Library Federation (DLF) and is maintained by the Library of Congress (2007)

[37] The latter migration typically entails migrating publisher provided Standard Generalized Mark-up Language (SGML) or Extensive Mark-up Language (XML) files to Portico's version of the National Library of Medicine's Archive and Interchange Document Type Definition (DTD).

content in the archive. Other formats received from publishers, PDF for example, are validated using JHOVE, where possible, before being preserved with their validation data in the archive. These other formats will be migrated in the future as acceptable archival structures for those formats are developed and adopted by the preservation community. Portico cannot predict when such migration efforts will be required, how much effort or expense they will incur, or how effective they will be. This approach is seen by some observers as naïve and short term, indicating a lack of serious investigation or experimentation. Nonetheless, the ongoing migrations, as described above, are routine parts of Portico's ingest operations.

On the one hand, despite its non-profit status, Portico's credibility is viewed with scepticism by a number of observers, particularly in the USA, because its funding model makes it potentially vulnerable to pressure from publishers and it lacks a track record in scholarly stewardship. On the other hand, it has a relatively positive reputation in the European library community.

4.3.4 CLOCKSS and LOCKSS

The LOCKSS Alliance and CLOCKSS are both outgrowths of the open-source LOCKSS software project at Stanford University, whose purpose was to show that replication can play a key role in the preservation of digital artefacts. About 25 publishers of commercial and open-access content participate in the LOCKSS programme and 100 institutions in more than 20 countries use the LOCKSS software (Kenney *et al* 2006). The LOCKSS Alliance (established in 2005) is a membership organisation for libraries built around the use of the LOCKSS software. The CLOCKSS initiative, established in 2006, is a group of six libraries and 12 publishers, working to create a 'dark archive' for e-journals. These related efforts will be referred to collectively here as LOCKSS, except where indicated.

LOCKSS was not intended to solve the full preservation problem,[38] rather to demonstrate the value of keeping multiple, replicated copies of important digital files at different sites which can back up and cross-validate each other's collections. Its strategy for dealing with format obsolescence is to rely on chains of open-source software (which it expects to be written by others[39]) which can translate content 'on the fly' via multiple intermediate formats if necessary, when obsolete formats are accessed. Its use of open-source software gives its users a degree of control that is not present in Portico or KB; this makes CLOCKSS quite attractive to research libraries. Because it does not supply such software itself but merely assumes that it will exist, LOCKSS does not view itself as a true preservation approach. It differs from Portico in this perception, but in fact Portico has not done as much as LOCKSS yet to preserve its holdings. LOCKSS is viewed by some observers as an option for perpetual access, not for preserving scholarly content in the face of technical obsolescence.

LOCKSS appears to offer an interesting alternative to Portico, based on a very different preservation model. Although neither LOCKSS nor Portico offer long-term preservation, the LOCKSS replication approach to short-term preservation and its open-source software

[38]However, it does claim to solve the problem of preserving web-published static content.

[39]For detailed arguments as to why this is expected to be the case, see Rosenthal *et al* (2005).

have made it attractive to many libraries and other institutions. According to at least one of its members, the CLOCKSS effort has not yet found an appropriate business model and is not an established entity, although it has ingested content from publishers that represents about 10% of their total content. Nevertheless, the market is interested in LOCKSS and many parties appear to be hedging their bets by entering into agreements with both Portico and LOCKSS.

4.4 Costs of archiving and preservation

The overall cost of archiving and preserving digital objects includes the administrative costs of developing and maintaining an appropriate set of preservation procedures and performing them over time, as well as the cost of providing whatever physical storage and computational resources may be required. Some – although not all – of these costs are likely to be independent of the technical preservation approach that is adopted. In addition, there are the costs of ongoing research in digital preservation, either through (inter)national projects or a local research agenda necessary to run and further develop the archiving and preservation service. Appendix C provides an overview of the different factors affecting the costs of preservation and their likely relevance in the future.

The costs of KB's e-Depot are difficult to delineate because part of the expenditures are embedded in the library's general overheads (e.g. housing, data network, administrative services, etc.). Nonetheless, KB claims that the annual costs of e-Depot are in the order of €4 million (see Table 5).

The total costs are expected to increase gradually with the ingestion of more digital items over the next years. As mentioned above, based on very crude assumptions, it is possible to estimate the average preservation cost per journal or item. In 2006, e-Depot had six million digital objects (mostly e-journal articles) from 3,500 ingested titles, which implies that the average costs per journal were between €1,200 and €1,500. The average costs per object amount to approximately €1. Considering the expectation that the number of ingested objects (nearly 10 million objects by the end of 2007) will grow faster than the total costs, the annual costs per object are expected to decrease. As can be calculated from Table 5, KB attributes 24 percent of these costs to the international component of e-Depot.[40]

[40] The costs attributed to the international e-Depot are based on the personnel costs for processing objects from international publishers, the employment of a programme manager for international e-Depot, the personnel costs of several board members and unit directors (strategy development and formulation, papers and presentations at conferences, etc.) and costs for consultancy services (Alliance for Permanent Access). With regard to maintenance costs and depreciation, KB has assumed that the ratio is 75 percent and 25 percent respectively for the national and international e-Depot.

Table 5. Estimated cost breakdown of KB's e-Depot, distinguishing the international and national e-Depot

	National[41]	International	Total
Staff costs:			
e-Depot Department	100	200	300
Digital Durability Department	350		350
International e-Depot management		80	80
IT Department	130		130
Management	5	40	45
Sub-total staff costs (€, 000s)	**585**	**320**	**905**
Material costs:			
Digital preservation Projects (Planets, Webarch)	1000	0	1000
Alliance	0	150	150
Maintenance contract	450	150	600
Annual depreciation	750	250	1000
Sub-total material costs (€, 000s)	**2200**	**550**	**2750**
Total costs (€, 000s)	**2785**	**870**	**3655**

Source: KB (2007)[42]

As Appendix C points out, on the one hand, the bulk of the costs of preservation procedures are expected to be attributed to repeated events. In other words, the average costs for an object in e-Depot are expected to rise if objects become technologically obsolete *and* if consequent preservation actions (such as migration) are required to deal with this obsolescence. On the other hand, if obsolescence does not require additional preservation action (as, for example, would be the case if emulation is used to enable obsolete objects to be viewed indefinitely), then the average cost per object may not rise. In contrast, due to economies of scale, the costs per item are expected to decrease with more items. Also, infrastructure, hardware and other material are expected to become less expensive.

Despite these uncertainties, the expected costs per title should be relatively insignificant vis-à-vis the revenues of a journal. When assuming that publishers alone will cover all of these costs, the contributions from large publishers (more than 170,000 objects)[43] will have

[41] The National e-Depot includes objects from: NTvG, members of the Dutch Publishers Association, CD-ROMs, DARE and the web archive. Note that national e-Depot also includes the publications of international publishers (e.g. Elsevier, Kluwer Academic Publishers, Brill and IOS Press) with 'The Netherlands' imprint. In Table 4, these have not been distinguished from the publications with other imprints however. It is difficult to distinguish between national and international imprints for these international publishers as they shift over time.

[42] Personal communication with Hans Jansen, Koninklijke Bibliotheek, 26 August 2007.

[43] We consider publishers with more than 100 titles to be 'large'. We use the number of titles as the basis for this because, thus far, the current number of titles is the best proxy for revenues (this may change with the increasing long-tail value of journal articles). In 2006, 3,500 ingested titles included around six million objects, which implies that on average a title has a total of around 1,700 articles. Thus publishers with more than 100 titles are assumed to have more than 170,000 objects. We consider publishers with fewer than 10 titles (around 17,000 articles) to be small. It must be emphasised that the distribution of numbers of articles per title is

to exceed €170,000 per annum, assuming €1 per object. The costs for small publishers (fewer than 17,000 objects) will be up to around €17,000 per annum. These costs will decrease with increasing opportunities of scale. Using the total number of ingested objects by publisher (as presented in Table 4), it is possible to calculate the total e-Depot cost breakdown per publisher.

Since KB argues that the costs of the national e-Depot are those incurred for its public role as a deposit library, the marginal costs of scaling this system to an international level are 31 percent (see Table 5). If publishers are expected to cover only the costs of the international e-Depot, annual fees could be 24 percent of those suggested above.

4.5 Trends and uncertainties in preservation

This section outlines a number of developments that can be observed in the field of digital preservation. While for some of these a clear trend can be distinguished, allowing predictions about the future, other developments are characterised by a high degree of uncertainty. Here we mention six relevant developments.

4.5.1 Demand for preservation of original inherently digital objects

Previous chapters have argued that increasingly, scholarly dissemination will include inherently digital artefacts, many of which may include dynamic components, such as executables and databases. Such inherently digital artefacts cannot be represented as page-images but instead must be rendered actively by computer programs executing on computers. This active rendering underlies the ability of such artefacts to exhibit their original dynamic – and potentially interactive – behaviour. However, as yet, research librarians, scholars and other users of digital repositories are unconvinced of the importance of preserving the original behaviour of such artefacts, at least in part because the scholarly record still consists largely of static page-image artefacts, whose behaviour is quite trivial.

This issue involves two significant and interrelated questions for preservation.

1. To what extent will inherently digital artefacts come to comprise a significant part of the future scholarly record?
2. To what degree will future scholars demand that such preserved inherently digital artefacts retain their original behaviour?

If the answer to both of these questions is 'not much', then preservation by means of migration may be sufficient for most scholarly purposes. However, if the answer to either question is 'to a significant extent', then migration (or any other approach that lacks the ability to retain the original behaviour of inherently digital artefacts) will prove insufficient to the task of ensuring the integrity – and thereby the authenticity – of the future scholarly record. Since few other institutions are pursuing emulation as a preservation strategy, KB

skewed, because there are many relatively young journals (with relatively small numbers of articles) and relatively few old ones (with a large number of articles).

could become a nearly-unique source for original digital objects, assuming that its emulation approach can be made to work.

> *'In order to stand on the shoulders of giants, one must know whose shoulders they are.'*
>
> Paul Courant, University of Michigan

The main relevance of this discussion for this report is that KB's multi-pronged approach to preservation provides it with a (so far) unique capability to preserve the behaviour of inherently digital artefacts, which may be crucial to ensuring the authenticity of the future scholarly record. Previous sections pointed out that often, the need for long-term preservation is not on the agenda in the stakeholder community. Stakeholders understand that there may be a problem with technical obsolescence on a long timescale, but there is a lack of incentives to commit them to collective solutions. However, libraries do express an explicit demand for an insurance policy to support perpetual access and large publishers recognise demand for agreements with safe places.

4.5.2 Access policies of preservation models

Observers have emphasised that the access policy of a digital repository is an important issue. The increased long-tail value of publications may lead to shifts in the business models of publishers, universities, information hubs and individual authors – any or all of which may seek to maximise their potential revenue streams by restricting access to their holdings via libraries and other information providers. Publishers may be uncomfortable with an on-site access regime, since these pose a risk of copyright violations by ill-intentioned visitors. Combined with copyright extension, this may complicate a digital archive's policies on providing online or even onsite access. The question is: will maintaining a 'dark archive' for digital preservation potentially conflict with the role of a (national) library? If publishers perceive increasing financial incentives to retain exclusive rights to their published materials for longer periods of time, KB may need to create suitable new agreements in order to ensure that it can maintain an up-to-date scholarly collection with appropriate access restrictions.

Similar concerns can be observed for access regimes after a trigger event. With the increased long-tail value of publications, ideally publishers would be looking for a digital archive which has mechanisms in place, in order to ensure that only those who have license rights for the digital content can access it after a trigger event. This would require a register of access rights, which would be very cumbersome.

4.5.3 A consumer market for preservation

Some observers think it possible that a consumer market for preservation services may arise, as individuals become increasingly concerned about preserving their personal email, digital music, photographs and video recordings, financial and medical records, documents, websites and programs. Similarly, organisations will need preservation services for their digital objects other than e-journals. Archiving demands are similar for such functions as for preserving national heritage. Also, comparable needs can be observed for archiving patient records, patents, financial accounts, pharmaceutical drug information, seismological data for the oil industry, digital music and film assets for the entertainment industry, meteorological data, human genome sequences, etc. Further, the inherent long-

term nature of business processes in the pensions and insurance sector requires a sustainable archive infrastructure in which digital preservation is a potential component. Every industry that needs to document and preserve information about its production processes for the future has a potential need for a digital archive. Preservation techniques that use platform independent emulation, such as DIAS, can be applied beyond the academic literature. If this happens, the scholarly community may benefit from the development of technologies and services in this broader market. A market for preservation that extends beyond the scholarly community could potentially arise. It is possible also that third parties in other industries which already provide some functionality for preserving their data will be able to assume a role in preservation of STM publications (interview with Reinier van Langen 2007).

It may take a while for these non-library parties to meet the stringent preservation requirements needed for authentic long-term scholarly preservation. Furthermore, the preservation requirements for these other sectors may be different because of copyright regimes, business confidentiality and privacy, etc. It is commonly assumed that less stringent preservation requirements (i.e. a lack of concern with the authentic preservation of original behaviour) would admit lower cost solutions. However, as has been noted, some techniques (such as emulation), which may be suitable for behaviour preservation, in fact appear to be less costly than others (such as migration), which are unable to preserve the behaviour of originals. Therefore, cost savings may have unexpected consequences, such as the adoption of more universal techniques (e.g. emulation).

Combining these other archiving and preservation needs in a digital archive for a wide range of materials could yield substantial economies of scale. If this is the case, it is not unlikely that a government would require convergence of techniques for reasons of public service value-for-money. In this case, the results of digital preservation research and experience would be applied beyond scholarship or publications. A legitimate question would then be: should such archiving be done by a national library? Regardless of the answer to this question, in the end, national libraries will benefit from the wider applicability of preservation technology and practice.

4.5.4 Government support for preservation

Because national libraries gradually are assuming a more proactive role in the digital environment, they feel that it is their obligation to retain a fairly comprehensive archive of digital knowledge. It is widely acknowledged by stakeholders that governments have a natural responsibility for preserving scholarly output for future generations. However, it is unclear what it means to be a national library in the digital, networked environment. National boundaries seem to become more and more irrelevant to networks and indeed to most scholars and researchers. As mentioned previously, national libraries – along with their government supervisors – will have to re-evaluate their policies continuously in order to decide what makes sense for them to collect, curate, preserve and make accessible in a networked world.

Government support for digital preservation varies considerably. The British Library, the National Library of New Zealand and the Deutsche Nationalbibliothek are embarking on ambitious (and costly) projects to preserve digital publications. Similarly, financial support from the Dutch Ministry of Education, Culture and Science for e-Depot has been

substantial in recent years (see Section 4.4). In other countries, both within and outside the West, support for such initiatives is not necessarily equally generous. Furthermore, some observers deem it not unthinkable that in the future, priority for preservation of scholarly publications may decline. This has happened in the past, when the future value of new technologies and their ability to perish were not recognised initially.[44] Furthermore, other more urgent societal priorities may arise in the future (for example, stemming from national, regional or global disasters) which may overshadow the relevance of digital preservation. In order to secure the sustainability of preservation initiatives, their funding resources should be distributed to reduce dependency on a single source.

> *'It is not unthinkable that in the future priority for preservation of international scholarly publications will reduce or disappear.'*
>
> Herman Bruggink, Nyenrode Business University
>
> *'It is uncertain whether the Dutch government will continue to financially support the e-Depot in the long-term future.'*
>
> Kurt de Belder, Leiden University
>
> *The e-Depot is a great gift from the Dutch government and the Dutch taxpayer to the international academic community. But can we expect the Dutch government to continue funding such archiving services of research material for the entire world into the indefinite future?'*
>
> Paul Ayris, University College London
>
> *'Governments are sometimes unpredictable paymasters. Unforeseen reasons may influence financing decisions.'*
>
> Cornelis van Bochove, Netherlands Ministry of Education, Culture and Science

Many observers acknowledged that in a digital world, where national boundaries are disappearing, the responsibility for preserving scholarly output could be placed at supranational level. They foresee a coordinating and facilitating role for the Commission to support European initiatives in this area. The European Commission already has an active agenda to support European cooperation in the field of digitisation, preservation and archiving (see for example, Reding 2007). Currently the European Commission is funding research on these subjects. This will continue and be reinforced with the 7th Framework Programme for Research and Development. In particular, under the Capacities Programme, the Commission will fund new developments to seed European e-infrastructures of digital repositories. However, it is not likely to fund operational activities, as this would require long time engagements that the Commission cannot commit to (interview with Carlos Ferreira Morais-Pires 2007).

[44] For example, websites from the early years of the Internet, and early television programmes, have been lost forever.

4.5.5 Government regulation

Government regulation will drive the future outlook of the preservation environment. While other regulations also may apply, we mention two policy areas.

First, to date, copyright regulation has had a strong and lasting impact on the dissemination and archiving of scholarly information. Copyright generally limits the ability of libraries to disseminate publications, thus putting a cap on their ability to generate revenues from their archival holdings. If copyright restrictions remain unchanged – and some argue that they will become even more restrictive (interview with Richard Boulderstone 2007) – publishers will have little incentive to deposit information in an archive that is publicly accessible, particularly with the continuing trend toward the increasing long-tail value of publications. Conversely, a more lenient copyright regime would allow open-access repositories to reduce their embargo time on copyright material.

Second, some observers (e.g. interview with Paul Ayris 2007) have mentioned the impact of European value-added tax (VAT) regimes on library practices. While books, newspapers and print periodicals are subject to a reduced rate of VAT in nearly every country of the EU 27, electronic publications are charged at the standard rate.[45] Obviously, this perverse incentive is hampering the ongoing shift from print to electronic publications. It is cheaper for libraries to order print versions of resources in addition to the electronic version, even if only the electronic version is considered valuable (Dewatripont *et al* 2007). Thus future changes in VAT regime may affect demand for digital preservation.

4.5.6 Technological uncertainties

Technological developments are among the most uncertain of all factors in this arena, since technology evolves so quickly and unpredictably. In particular, digital formats, which are the technical core of the preservation problem, may evolve in unknown ways as new formats are developed to provide new, unforeseen capabilities. In addition, the popularity of proprietary formats – of which there are many – tend to ebb and flow with the fortunes of the companies that own them. The ICT vendor market is still potentially quite volatile, despite the presence of large, seemingly permanent companies such as Microsoft, Adobe and Oracle, and new vendors with new formats may still appear with little notice. Finally, future paradigm shifts – which often follow new ICT hardware developments – may lead to novel formats, forms of scholarly communication, research, publication and usage patterns and preservation challenges. Other examples of technological trends are discussed in Appendix C.

4.5.7 Coping with uncertainty

The above discussion of the current state of digital preservation and archiving has described the major forces, stakeholders, trends and uncertainties which are likely to affect the evolution of this dynamic endeavour in both the immediate and more distant future. Although the situation involves a great deal of uncertainty, we believe that this uncertainty

[45] Only in Hungary are both rates equal (20 percent), whereas seven Member States' rates have multiple VAT levels for print journals, the highest one being equal to the e-Journal VAT level. For example: UK has 20 percent for e-journals and 0 percent for print; The Netherlands has 19.6 percent for e-Journals and 6 percent for print. See also European Commission (2007: 15).

can be categorised and understood in such a way as to help KB develop a strategy that will be relatively robust in the face of future developments. The next chapter attempts to analyse this uncertainty in terms of the forces that are driving it. It then develops a small set of scenarios to explore the combined effects of these drivers.

CHAPTER 5 The uncertain future of preserving the past

The preceding chapters describe a fluctuating environment in which it is very difficult to make firm predictions. Yet KB needs to develop a strategy for digital archiving and preservation that is robust in the face of these uncertainties. In this chapter, we analyse the driving forces that underlie these uncertainties and develop a set of scenarios which can help KB to anticipate and prepare for future developments.

Scenarios are analytical tools that are used to represent and deal with uncertainties.[46] Each scenario is a description of one possible future state of the system. Scenarios do not forecast what will happen in the future; rather they indicate what can happen. Also, scenarios do not include complete descriptions of the future system; they include only factors which might strongly affect the outcomes of interest. Because the only sure thing about a future scenario is that it will not be exactly what happens, several scenarios, spanning a range of developments, are constructed to span a range of futures of interest. No probabilities are attached to the futures represented by each of the scenarios. They have a qualitative, not a quantitative function. Scenarios do not tell us what *will* happen in the future; rather they tell us what *can* (plausibly) happen. They are used by decision-makers in both the private and public sectors to prepare for the future: to identify possible future problems and robust policies for dealing with them.

While the scope of this study does not allow for a fully-fledged scenario-planning exercise, we have developed examples of possible scenarios for the future of digital archiving and preservation. The four scenario examples resulting from this chapter should be interpreted as an illustration of how the future may develop, what the impacts for digital preservation may be, and how this may influence KB's position and strategy.

5.1 Scenario development

A scenario can be defined as a consistent and plausible picture of a possible future reality that informs the main issues of the strategic policy debate. A scenario is a static presentation of the future in which only the future reality itself matters, not the way in which one gets there from the present.

[46] For more information on scenario planning, see for example, Deweerd (1973), Kahn and Weiner (1967), Kahn *et al* (1976), Schwartz (1991), Smith (1964).

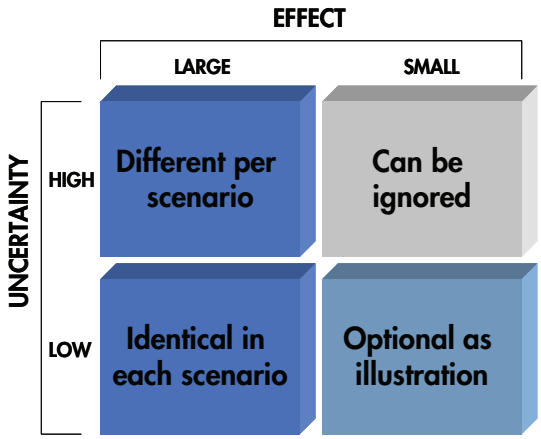

Figure 9. The building blocks of scenarios

The use of the term 'scenario' as an analytical tool dates from the late 1950s, when researchers at the RAND Corporation defined states of the world within which alternative weapons systems or military strategies would have to perform.[47] Since then, scenario planning has been employed by many successful organisations and enterprises of all sizes.

Building scenarios is an exercise in *discipline* and *creativity*. The discipline is needed to structure the set of scenarios so that they reflect the issues requiring exploration. Creativity is needed in filling out the scenarios so that they become meaningful, consistent and plausible.

RAND has developed a six-step approach to the development of scenarios which has been used in many other projects (see for example, RAND Europe (1997). The approach includes the following six generic steps.

Step 1: Specify the system and define the outcomes of interest.

Step 2: Identify external factors driving changes in the system.

Step 3: Identify system changes, connections between these factors and system changes and how the changes affect the outcomes of interest.

Step 4: Categorise the uncertainty of the factors and system changes.

Step 5: Assess the relevance of the uncertain factors and system changes.

Step 6: Develop scenarios using the factors according to the combinations in Figure 9.

The dimensions to consider are constructed principally among the high uncertainty items with large effects. This is also where the attributes within each dimension to consider are established and elements that represent the attributes are selected. The number of scenarios chosen will be a reflection of the space of strategic policy issues under consideration.

[47] See for example, Deweerd (1973), Kahn and Weiner (1967) Kahn *et al* (1976), Smith (1964).

5.2 Driving forces of digital preservation

From the developments identified in Chapter 3 and Chapter 4 we can identify several drivers which are characterised by a high degree of uncertainty and potentially have the greatest impact on the future outlook of preservation.

5.2.1 Methods of disseminating scholarly information

Chapter 3 examined the uncertainties surrounding future scholarly dissemination and publishing. These concern the roles and strategies of institutions, including publishers, universities, university libraries, national libraries, individual authors and information hubs. One such uncertainty concerns the possible transformation of the traditional peer-review process in response to the widening bifurcation of the publication stream into formal and informal branches. If this leads to the increased use of informal publication as a source of review, then such informal material may strengthen its bid to become part of the scholarly record. Similarly, the possible atrophy of the university tenure system (of which there are a number of signs) may undermine the traditional peer-review process, which might lead to the increased importance of informal publication. Either of these trends could encourage KB to include more diverse forms of informal scholarly publication in its repository, in order to preserve the scholarly record. Another issue concerns the possibility of university libraries taking on the roles of institutional publishers and/or institutional repositories: this could conflict with KB's vision, although it might offer synergy with that vision.

As mentioned previously, with journal publication being the prime channel of scholarly dissemination, it has been a remarkably robust and effective mechanism. However, there has been a gradual increase in the price of subscriptions, leading to journal cancellations and indirectly to a slow shift towards alternative channels of dissemination. (Examples of these alternative channels have been discussed in Chapter 3) While currently the impact of these alternative channels is very limited and it may vary by discipline, a trend break is not inconceivable: the traditional publishing model may or may not continue to dominate in the STM sector, although it seems likely to retain a significant share of that sector for some time.

In parallel, information hubs, particularly those that include search engines such as Google and Yahoo!, offer technologically sophisticated routes to information. These information hubs have some important drawbacks despite scholars' interest in them. For example, there is no indication of how complete the search results are, and there is no commonly-accepted quality assurance mechanism to evaluate what is found. However, undoubtedly the sophistication of search algorithms will continue to improve and alternative quality standards may arise. Search processes are becoming increasingly important as the volume of information disseminated and stored continues to escalate. The ascendence of search engines represents a new form of dissemination that is already widely perceived to have eclipsed traditional channels for most students. Although these engines do not generate information, they can be very effective at 'publishing' it (i.e. making it public) without recourse to traditional publication, marketing or distribution mechanisms. Furthermore, the information hubs that maintain these search engines are attempting to position themselves as owners and purveyors of information, which in turn will have to be preserved. It remains unclear what future role these hubs may play with respect to scholarly

STM information and its preservation; precisely for this reason, their behaviour should be monitored and analysed as it evolves.

These new means of scholarly dissemination could make it easier for scholars to by-pass traditional publishers. Some segments of the scientific community, for example high-energy physics, already are operating partly outside the publishing and library loop. The extent to which scholars will shift to more informal dissemination channels is uncertain. However, the consequences for preservation are important. Scholars' increasing reliance on informal sources and on the data and models underlying formal publications will force a digital archive to expand its holdings beyond the output of traditional publishers. In an extreme case, the number of actors involved in disseminating scholarly output may become so vast that it is infeasible to approach all of them for archiving. Increased reliance on information hubs and other 'pull' mechanisms may gradually make KB, and libraries in general, secondary rather than primary sources of information for scholars, unless they can ally themselves safely with such information hubs or provide similar searching and alerting services themselves. Nevertheless, regardless of whether or to what extent these informal sources usurp the role of libraries as primary information providers, the information that they maintain will still require preservation, yet they may be ill-equipped or disinclined to provide archival services. Therefore, preservation providers such as KB may need to develop agreements with such informal services in order to ensure that a complete scholarly record is preserved.

5.2.2 Digital archiving and preservation of scholarly information

Many industry thinkers regard the discussion about the business model (i.e. 'who provides access to whom, in what form and at which moment and who pays') as the most difficult one that publishers, libraries and other stakeholders face. In this section we discuss three uncertain drivers of future business models.

Firstly, a digital archive can have one or more distinct functions (see Table 6). It is possible that these functions can be fulfilled by a single holistic service that takes account of all requirements for these different functions. In particular, KB aims to provide (on-site) scholarly access and perpetual access service and to preserve cultural heritage. However, the requirements of research librarians, publishers and scholars for digital archiving and preservation overlap in a number of ways, including their need for authenticity.

Table 6. Different rationales for digital archives

Perspective	Main function	Audience that benefits	Form of content required
Scholarly access	Facilitating access to current intellectual content and insurance policy for perpetual access	Research community	Easy online access to objects retaining as much of their original behaviour as possible, plus vernacular renditions for more casual use
Cultural heritage	Preserve the historical context and cultural meaning of important artefacts	Future generations of users, historians, general public, governments and non-profit institutes	Vernacular renditions to make the core contents of originals more easily accessible and comprehensible, plus authentic originals to serve as venerated historical artefacts
Economic value	Facilitating access to economically-profitable content	Publishers	Unmodified (un-republished) originals plus repackaged or repurposed versions

The need for behaviour preservation

A significant challenge facing archiving and preservation approaches is whether and how to retain the original behaviour of preserved artefacts. Some parties believe that preserving the information content of text, numbers and images is crucial, but that formatting and fonts, etc. are not vital.[48] Others, including KB, believe that the authenticity of text, format and medium are all required for comparative purposes, because it is impossible to predict which elements of an artefact may be important in the future; thus they regard preservation of either the original or something very close to the original as essential. Furthermore, as argued above, inherently digital artefacts exhibit behaviour that goes far beyond that of static page-image documents, including dynamic, multimedia, executable and interactive aspects which cannot be preserved at all by simply extracting their textual or visual content. KB's specification for the DIAS e-Depot system and use of emulation techniques reflect this concern. A third group, which includes Portico, considers the transformation of a document from one presentation form to another to be a technicality that will be resolved once the software packages currently in use expire and are no longer operable, although this ignores the issue of inherently digital behaviour.

Traditionally, national libraries have emphasised the importance of long-term preservation of content in its original form, as this retains as many attributes of the original artefact as possible, thereby serving the broadest range of possible future uses while simultaneously maximising the authenticity of the artefact. This authenticity is of great value to scholars; even though they often utilise less authentic vernacular renditions of artefacts for the less critical aspects of their research, the original is always the ultimate arbiter of truth.

As yet, the scholarly community is not widely concerned about the authenticity of preserved digital artefacts, being concerned mostly with perpetual access of their intellectual content. However, it seems likely that this lack of concern stems from a combination of two factors:

1. most digital artefacts are still page-images, whose preservation is relatively straightforward, rather than inherently digital objects whose meaning (and even existence) depends on their executable behaviour; and
2. multiple conversions of digital artefacts into successive formats has not been necessary yet over the short lifespan of most such artefacts, so that the cumulative corruption that is likely to result from such conversion has not infected significant numbers of artefacts yet.

In addition, as indicated in the third column of the table, scholars are not the only users who may be concerned with preserving authentic originals. Historians, the general public, governments and non-profit institutes may require such authenticity, whether to recreate historical perspective, enforce legal and ethical accountability or simply to venerate important artefacts. For all of these purposes, it may be necessary to retain the original

[48] As a counter example, consider the 1999 scandal involving French finance minister Dominique Strauss-Kahn, which was uncovered in part due to the fact that a backdated digital letter that he created to negate the charges was found later to have used a character font which was not developed until after the purported date when the letter was written. See for example, Vinocur (1999).

behaviour (as well as the look and feel) of historical documents, in addition to their core intellectual content.

Even publishers may have a stake in the behaviour preservation of their original publications, by analogy to the value of first editions in the print medium. KB has interpreted its agreements not to republish the works of publishers to be a requirement that digital deposit holdings not be reformatted or modified in any way; this requires the authentic preservation of digital originals. Publishers may want to reformat, repackage or repurpose their publications themselves in order to create new value streams (the long-tail), in which case they may be particularly concerned that libraries and other repositories preserve their materials only in their original form, so as not to compete with them.

Furthermore (as we have emphasised repeatedly), inherently digital artefacts are impossible to preserve other than by preserving their original behaviour. The core intellectual content of such an artefact generally does not exist except as a manifestation of its behaviour, which goes far beyond that of static page-image documents, including dynamic, multimedia, executable and interactive aspects which cannot be preserved at all by simply extracting their textual or visual content. In some cases it may be possible to extract vernacular renditions of such artefacts in contemporary future forms, but there is no guarantee that these renditions will be capable of retaining the artefact's full original behaviour and meaning. Even if they do, the original must be preserved in order to generate these vernacular renditions from an authentic version (rather than a previous vernacular rendition) in the first place.

These factors argue for the development of mechanisms to preserve the authentic original behaviour of inherently digital artefacts, whether or not page-image artefacts can be preserved adequately for most purposes by other means. As noted elsewhere in this report, both logic and the available empirical evidence suggest that preservation of original behaviour may be more affordable than static preservation (for example using migration), even over relatively short timeframes.

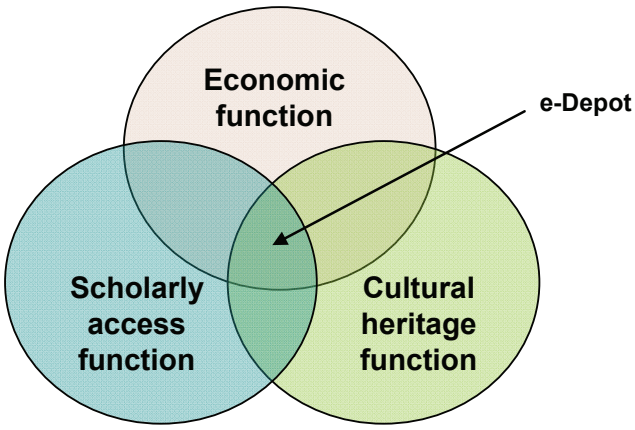

Figure 10. Possible compatibility of the different functions of digital archives

As illustrated in Figure 10, KB's strategy assumes that scholarly access, cultural heritage and economic functions are complementary. The above analysis reveals that all three of these functions share some need for the authentic preservation of original behaviour,

especially for inherently digital artefacts, making it unlikely that the overlap shown in the figure will disappear. Nevertheless, it seems to be the case that scholars, publishers and other users are relatively unconcerned as yet about the authentic preservation of inherently digital artefacts. Therefore, it may be difficult for KB (or any other organisation) to sell a business model based in part on the ability to preserve such artefacts. The degree to which the market for preservation may evolve to value such behaviour preservation remains a key source of uncertainty.

The possible rise of a consumer market for preservation

A second uncertain development discussed in 0 is the extent to which a consumer market for preservation may arise. Will preservation and archiving functions of patient records, financial accounts and human genome sequences, etc. be combined with the preservation function of scholarly communication? Whether this will happen depends on a number of underlying drivers, one of which is the political impetus for more accountable public services. We have seen developments of similar deregulation, privatisation and (international) consolidation in other sectors (e.g. telecommunications and energy supply). A government could push for synergies (and possibly competition) with other archiving and preservation functions, which could lead to integrated archiving services that go beyond scholarly communication.

Public support for digital preservation

A final factor characterised by high uncertainty and high impact for the future of preservation concerns future public support for digital preservation. Under normal circumstances, public funds are expected to continue supporting the preservation of scholarly output for future generations. However, there are plausible 'wild cards'[49] imaginable that will cause a trend break in the priority for digital preservation, or preserving cultural heritage in general.

5.3 Possible future scenarios for dissemination and preservation

The outlook for archiving and preservation that emerges from the preceding chapters is complex. The previous sections have illustrated that this outlook is dependent partly on the outcomes of uncertain developments in scholarly dissemination and publishing. It will depend also on the attitudes that prevail regarding archiving and preservation itself. Scenario planning is a practical approach to facilitate the process of anticipating the span of plausible futures.

In this study we have constructed four scenarios by selecting two driving forces (as described in Section 5.1). The possible future development of these drivers can be explained by a spectrum of plausible states. By focusing on the extremes of this spectrum,

[49] In decision theory, wild cards refer to low-probability, high-impact events. Such sudden and unique incidents might constitute turning points in the evolution of a certain trend or system. See for example: Petersen (2000) or Van Notten *et al* (2005). In the current context, such wild cards might include major shifts in priorities following political, economic or natural disasters.

we can construct four possible scenarios that span the range of plausible futures. These drivers are:

- scholarly communication and publishing – bifurcation versus consolidation;
- digital archiving and preservation – low versus high priority for the authentic preservation of inherently digital originals.

First, scholarly information can vary from, at the one end, further consolidation and domination by the large traditional publishing houses, to the other extreme, in which scholarly dissemination is dominated by the new, informal dissemination techniques discussed above in Section 5.2.1. Second, archiving and preservation can be driven by the dominance of concerns about perpetual access (by research libraries) and technical obsolescence (by governments) that lead to a demand for more sophisticated and better-organised approaches, as well as investment in the development of alternatives. However, it is possible that relaxed attitudes regarding these long-term uncertainties may prevail, in which case more ad hoc and short-term archiving solutions may prevail.

These two dimensions combine into the four scenarios outlined in Figure 11. The grey area in the figure illustrates the current situation, in which traditional publishers largely dominate the world of scholarly dissemination, and attitudes vary across a wide range from serious efforts to assure permanent access and avoid technological obsolescence, to a focus on developing business models which have the right economics, regardless of access and technology concerns. The resulting four scenarios are described below.

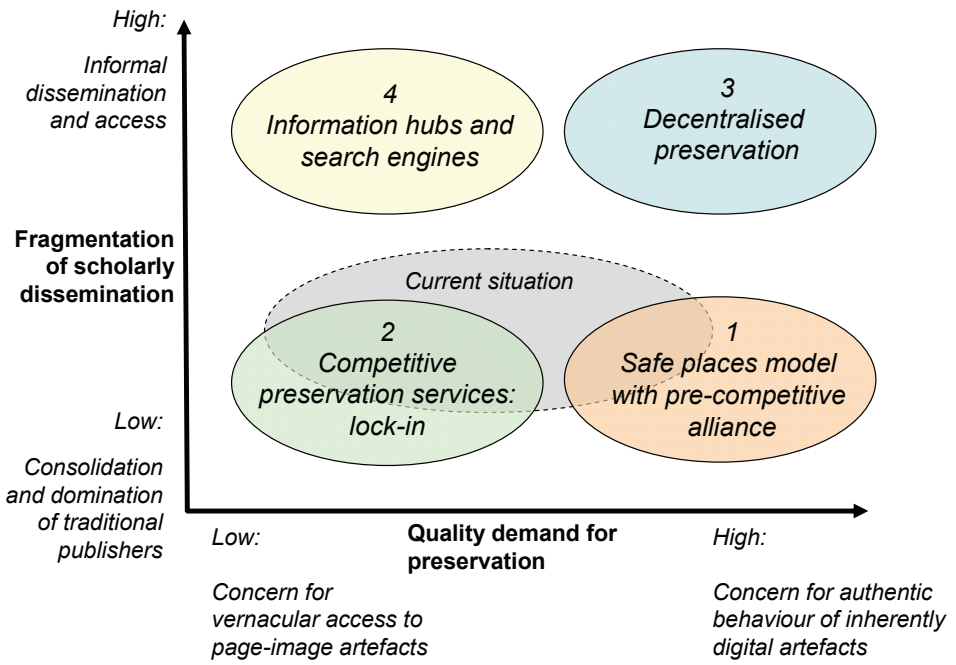

Figure 11. Scenario framework for the future of digital archiving and preservation

5.3.1 Scenario 1: global safe places model (pre-competitive alliance of preservation initiatives)

In scenario 1, which is an evolution of the current situation, the major publishing houses have continued to dominate scholarly dissemination and the industry is characterised by further consolidation. Publishers consider archiving to be an activity that on the one hand,

is non-core, but on the other hand, is essential to their reputation for reliability. Due to high concern for the preservation of authentic behaviour, there is sustainable government funding and funding from charities available. A convergence of different preservation approaches leads to one or two *de facto* standards. A limited set of about six geographically safe places is considered to be sufficient for the challenging and technologically complex task of long-term preservation. Archiving and preservation becomes the domain of a small number of reputed expert organisations. Technological uncertainties may force publishers to hedge their bets and sign archiving agreements with multiple, geographically-distributed institutions. Furthermore, there is a push from policymakers and funders to collaborate and seek alliances. This occurs for several reasons: to avoid unnecessary redundancies and yield efficiencies; to ensure (worldwide) completeness of the preserved corpus; and to foster international sharing of expertise. This pre-competitive alliance has introduced a registry for the worldwide preservation of e-journals with designated safe places.

5.3.2 Scenario 2: more competitive environment (lock-in of preservation services)

In scenario 2, publishing houses have continued to dominate scholarly dissemination. However, there is little concern about long-term authentic digital preservation. Governments have reduced budgets for digital preservation initiatives drastically and urged those to be self-sustainable. Consequently, the situation becomes more fluid, as publishers adopt a more business-like attitude and opt to support the drive to develop economical business models. This may lead to the entry of new players in the archiving and preservation business and an unbundling of the traditional players. The safe places with robust funding mechanisms have incentives to seek different cost-effective solutions to their preservation challenges. Research libraries sign a perpetual access agreement with only one safe place that is inexpensive and meets their requirements, just as if they were selecting car insurance. Since there is little concern for technical obsolescence or authenticity, sophisticated preservation technologies may be dropped in favour of less sophisticated approaches.[50] Thus there is an opportunity to lock in relationships through active acquisition. Safe places need to advertise their systems in competition with other safe places.

5.3.3 Scenario 3: decentralised preservation

In scenario 3, traditional publishers have lost ground to the open-access movement. The research community has found and accepted alternatives for peer review which traditionally had been facilitated by publishers. Now there are mechanisms in place that ensure the quality of e-papers published in research institutions' own digital repositories, or by the authors themselves. The demand for centralised safe places providing a perpetual access insurance service is low, since publishers' monopoly position has disappeared, making centralised preservation less feasible, or at least less likely to be complete. At the same time, the awareness of authenticity issues is widespread. Therefore, authors, research funders and facilitators of information services (virtual libraries) will demand robust preservation mechanisms from their dissemination channels. Governments fund their

[50] However, we do not assume that sophisticated techniques such as emulation are either more complex or expensive than less sophisticated ones. In fact, migration could be the most complex and expensive, and therefore the most likely technique to be dropped under this scenario.

universities to develop decentralised preservation approaches that fit their needs, standards and economics. The prevailing concerns for the authenticity and integrity of objects will motivate people to demand technically-sophisticated solutions, for example, by seeking the advice of specialised consultants. Off-the-shelf emulation packages become available, allowing each institutional repository to implement digital preservation systems.

5.3.4 Scenario 4: Information hubs and search engines

Scenario 4 represents an arena that is in flux on all sides. This situation is subject to rapid change and is not very transparent. Scholarly dissemination has been bifurcated and largely transferred to decentralised scholar-driven channels such as open access and grey publication in institutional repositories, or by authors themselves. Information hubs and powerful search mechanisms play an important role in access to scholarly communication. The demand for centralised safe places providing a perpetual access insurance service is low because there is as yet little priority for behaviour preservation. Also, financial support for long-term preservation from government or charity sources is weak in this scenario. Dissemination channels, such as blogs or wiki-type hubs, are thriving and form an organic mechanism of knowledge accumulation. These dissemination channels are very difficult to archive and preserve, but they have tools for bitstream validation and integrity control. In addition, an important preservation mechanism is the facilitation of continuous access; content hosts are notified as soon as users experience disruptions.

5.4 Conclusion

This chapter showed that the future for digital preservation is highly uncertain, due to developments in scholarly dissemination and publishing and the unknown priorities of preservation stakeholders. Scenario planning is a useful tool to facilitate robust planning to cope with such uncertainties. However, developing scenarios is not enough. It is up to KB and the wider stakeholder community to test their preservation strategy in different unlikely but plausible scenarios. Section 6.8 provides first pointers to such a planning process.

CHAPTER 6 Strategy implications for KB

The previous chapters highlighted the fact that the world of scientific publishing, preservation and scholarly publication is facing a number of uncertainties. KB's relationship with publishers (as well as with other libraries and preservationists) makes it part of what may appear to be a self-contained publisher/preservationist dyad. This dyad will evolve on its own, in response to new technological and business model developments in publishing and preservation. The outlook for the future from within this dyad is relatively uniform and unconcerned with the potential for disruptive change which could come from developments in scholarly communications, or from the emerging role of new players, such as search engines. The persisting STM publication and business models remain at the heart of this outlook, which is generally in line with KB's strategic view – i.e. that a global network of a handful of trusted repositories should ensure the preservation of all STM publications. A natural role for national libraries in this area is generally accepted.

In order to assess KB's e-Depot strategy, first we will discuss what the continuing justification could be for the existing strategy, and the issues that KB needs to address in order to ensure the effective execution of that strategy. Our initial analysis assumes an evolving world, in which the main actors and the main objects of preservation nevertheless remain largely unchanged. Then at the end of the chapter, we factor in some of the more disruptive developments in scholarly communication and the underlying uncertainties that potentially undermine traditional publishing and preservation models, which may force publishers and libraries to adapt their strategies or to adopt new roles. In many cases, our analysis and conclusions support existing plans and policies of the KB, so they should not be construed as criticizing these plans and policies but rather as reinforcing and substantiating them.

6.1 The continuing justification of KB's activities

The role of a national library has to be adapted as a consequence of moving into the digital era. Although for some existing and emerging issues the consequences for a national library are clear cut, for other issues, the situation and the role of this public institution is more ambiguous. Here we emphasise four current or potential aspects of KB's remit that may require reconsideration.

6.1.1 Obligation to archive national STM imprint

KB's traditional public role – to store and preserve a copy of all (STM) publications whose imprint is 'The Netherlands' – is typical of national libraries. As discussed above, the

advent of digital publications, as well as the international nature of STM publications and their publishers, have created new challenges, risks and opportunities for KB and its peers. KB has evolved naturally along with the reality of internationalisation of scientific e-journals, thus moving beyond the original scope of its national remit. Although one can debate whether it is the task of a national library to preserve international publications at the cost of taxpayers' money to the benefit of the global scientific community as a whole and that of large STM publishers in particular, it is clearly in the overall public interest to do so.

Thus far, the Dutch government has supported this trend towards serving a wider interest than a strictly national one. However, there may be reasons for KB to ask the users of this service to shoulder some of the cost of preservation in the future. (The issues surrounding the financial base for e-Depot will be discussed in more detail in Section 6.4.)

6.1.2 Guaranteeing authenticity of scholarly record

Scholarly disciplines and society as a whole have a crucial need to maintain an authentic scholarly record for intellectual, scientific, pragmatic, legal, historical and ethical purposes. In order to guarantee the authenticity of this record, several criteria must be met. Preservation mechanisms must ensure that the record is not corrupted or lost inadvertently over time. In addition, security measures must ensure that the record is not intentionally modified for monetary, ideological or political purposes. Finally, verification mechanisms must be put in place to confirm that preservation and security procedures have been applied consistently and effectively.

In order to serve as a trusted steward for the scholarly record, an institution must be as far above suspicion as possible. This argues that preservation should be entrusted to publicly-funded or non-profit institutions. Furthermore, to guard against accidents, inadvertent corruption of the record or governmental tampering, a number of such institutions in different countries should have a (partially) duplicate collection and enable cross-monitoring of each other's collections. This is not only an important justification for continuing KB's archiving and preservation role; it also supports the thesis that an international network of repositories is needed.

6.1.3 Guaranteeing perpetual preservation

Outsourcing collections is very risky. These are a library's core competence or asset and should not be trusted lightly to others. Commercial parties, such as publishers, have incentives not to preserve for perpetuity, since the cost of doing so is likely to increase as the returns diminish. Similarly, there is no economic motivation for preserving publications that are not in demand.

This may change, however, if – as the result of pressure from scholarly communication or in the hope of finding new revenue streams – publishers shift their focus to the long-tail value of their assets, as ICT enables the repeated repackaging, reuse, mining and re-purposing of previously published material. Combined with extended copyright protection, these new capabilities seem likely to give publishers an increased stake in the longer-term preservation of their assets, while simultaneously making them more wary of relinquishing control over access to those assets via libraries or other mechanisms – at least to the extent that such access may decrease their potential for generating future revenue.

Depending on the potential margins that can be generated by the long-tail and the technical entry barriers to digital preservation, publishers may start (or in some cases extend) their own preservation activities or become willing to pay more for outsourcing such preservation services. In the latter scenario, it is likely that access provisions will be minimised.

Some of these developments seem to drive Google's digitisation project. Information and publications that do not seem valuable now, may be very valuable in the future, once information sets are complete and when there will be new technologies and applications that allow the combining of information sources to create new knowledge. Also, Google is effectively taking on a role of providing access to archives; however, if it is to offer guaranteed access in perpetuity, Google will have to provide effective, efficient, reliable and sustainable preservation services, although there is as yet little indication that it intends to do so. Libraries have concluded – and publishers believe – that they cannot rely on a disaggregated, non-organised system to achieve such perpetual access.

Thus, as neither search engines nor publishers will provide the guarantee of structured, perpetual preservation of entire collections, it is likely that a public service function remains to be fulfilled by national libraries such as KB.

6.1.4 KB's leading technology and expertise

KB has been a front-runner in the field of digital preservation and has been open in its ambition to share its knowledge and infrastructure (mostly for free). This role is not so much a formal one based on a legal requirement, but is borne out of KB's public service duty and its *de facto* leading knowledge position. The fact that it is a centre of excellence and an acknowledged innovator in its field is valuable in its own right. In many ways, digital preservation is still in its infancy and its importance is not fully understood, even among many key stakeholders. Thus there may be continued interest from government to invest in knowledge development in this area. In addition, a possible, though not a necessary, evolution for KB may be to exploit its knowledge through consultancy services and by offering its capacity to store materials and collections for others. The strength of this argument depends on a thorough cost–benefit analysis, in order to justify continuous investment in research and new capabilities and to understand the market value of KB's services.

For KB to continue these public service functions and allow it to roll-out its international strategy fully, it must take account of a number of issues, which have emerged from our literature review and, more importantly, from our stakeholder interviews. These are discussed in the following sections.

6.2 Awareness and trust in the stakeholder community

To become a trusted digital archive there needs to be trust from users – mainly libraries and publishers, but also scientists and scholars – in KB's independence, as well as its technological and managerial capabilities to deliver the expected quality of service in perpetuity, without external interference or manipulation. University libraries in particular have expressed this need.

There is no *a priori* reputational advantage in being a national library or representing a small country in a non-competing linguistic area. University libraries – especially in the Anglo-Saxon world – tend to focus on that world when looking for a trusted ally. Publishers seem less concerned when choosing archiving capability and are more driven by concerns about price, reputation and the risk of intellectual property rights breaches due to libraries' public access function. Reputation is important, since a preservation service is offered to licensees (university libraries) as insurance for perpetual availability. Such insurance is valuable only if the preserving organisation is reputable.

The actual capability of a preservation service provider to deliver perpetual service does not appear to be of much concern within universities, since they have a general lack of confidence in current digital preservation arrangements. This appears to be driven at least in part by perceptions, since most stakeholders display limited knowledge of the intricacies of preservation, and neither publishers nor libraries seem to allow preservation quality concerns to guide their decisions. Other factors seem to have a greater effect on the perceptions of several stakeholders: for example, the notion that an insurance service that is provided for free cannot be much good. Moreover, interviewees representing international university libraries expressed doubts about KB's service delivery capabilities, since formal contractual guarantees between KB and university libraries are currently limited to The Netherlands.[51]

Although there is some scepticism about KB's remit, its reputation for preservation expertise is well known among Dutch university libraries, the large international scientific publishers and a select group of other (inter)national libraries. Beyond these, KB's reputation is less known. If KB intends to play a role as one of a small number of international safe places, it should market itself more globally as one of those guarantors of preservation and long-term access. Overall, KB is not yet widely regarded as having such a profile in North America, or even in Europe, despite its having actively contributed to a number of European research programmes and collective initiatives with other libraries.[52]

> *'e-Depot is still too unknown in The Netherlands and abroad. It should first establish a solid argument why Dutch taxpayers should continue to pay for e-Depot's international strategy and its ongoing R&D effort. Public support is an important guarantee for public funding.'*
>
> Cornelis van Bochove, Netherlands Ministry of Education, Culture and Science

Some universities and publishers express doubts about KB's ability to provide the required services in case of a trigger event. There is the feeling that the capacity is lacking to turn a 'dim archive' into a live resource. Existing contractual agreements with the publishers about reimbursement of costs and the conditions under which access may be provided do not seem to change this perception. This is especially the case with many university libraries, which indicate a lack of awareness of KB's exact mandate and how KB would

[51] Formal archiving agreements between the KB and the DARE repositories.

[52] It is interesting to note that the selected group of experts from US university libraries did know of KB and regarded it as further advanced in digital preservation, and in some ways incomparable to Portico and (C)LOCKSS, whereas in Europe the contrary seemed to be true.

handle the logistics of providing such access. Again, the fact that the services are provided for free reinforces this scepticism.

This lack of awareness among stakeholders is a potential barrier to KB's successful role as a leading safe place for digital archiving and preservation and its drive to form an international network of such archives, especially among university libraries that present the main demand for preservation – as they request perpetual access from publishers in their licensing contracts. The fact that KB is a national library and that it is not from an Anglo-Saxon country raises doubts about its intentions (although more so in Europe than in the USA). It appears that KB has not distinguished itself sufficiently from other preservation initiatives or national libraries, based on its preservation excellence. It should target universities more effectively, champion the case of preservation and publicize the risks of current (non-emulation) solutions, such as those offered by Portico, CLOCKSS and LOCKSS. The European Commission is looking for support in its digitisation and preservation initiatives under the i2010 programme. KB should maintain and possibly improve its position further as a leader in this field, and communicate its role within projects such as the European Digital Library, through existing networks for libraries, publishers and scientists.

6.3 Justify KB's national and international remit

As with meteorological, seismological, astronomical and most other research data, there will always be an element of public responsibility for the preservation of scholarly communication. There is also a clear and natural responsibility for a national library, which has the traditional role of conservator of national heritage. However, in the digital era, national imprint has become increasingly difficult to define and irrelevant as a selection criterion for setting the mandate of national libraries. This is a pertinent problem, since traditionally a national library's responsibility only covers national imprint.

There is no international governance structure that provides a formal remit or ensures a similar level of storage and preservation at the international level. Thus there is a possibility of significant overlap between national activities – especially for international publications (which nowadays are published mostly in native English-speaking areas in the West) – and, more importantly, of possible gaps in the preservation coverage of global STM production. This concern may lead KB to feel a duty to take up an international responsibility, though at the moment there is no formal mandate by any public authority to do so.

At the European level, the issue is being followed with great interest. However, the European Commission is deliberately wary of getting involved with STM publication and preservation due to its sectoral, organisational and geographical specifics across Europe. Instead, it has focused on connecting knowledge centres (mainly through the GÉANT network) and digitising and preserving European cultural heritage. STM preservation is seen as an important issue, and support is given to digital preservation research initiatives to cooperate and share best practice. The Commission is looking for bottom-up initiatives and natural leadership from the current centres of excellence and the most important national libraries involved in digital preservation.

As such, KB continues to position itself as a leading expert in preservation and helps to take the lead in the creation of a European (or global) network of national libraries and STM repositories.[53] However, it cannot be expected that there will be a formal mandate at EU level to develop a European digital STM library or network. Also, there is some scepticism among KB's peers as to its intentions ('Why would a small country's national library have the ambition to become the STM archive for Europe?'), and they may want to prevent KB from developing such a leadership role, thus pursuing their own developments instead of cooperating with KB. Some of this hesitation to cooperate and share may be driven by cultural and historical developments, which are largely determined by the national remit of these institutions and their pride in maintaining their own collections. In addition, they may aspire to become leaders in the field themselves, as this is likely to bring benefits to the institution and its country.

Herein also lies the additional argument for building the exemplary role of e-Depot, as it is likely to create an environment of innovation and development which also will attract further technology and commercial investment to The Netherlands. Undoubtedly, there will be a number of synergistic effects on other scientific fields, and the knowledge generated in the digital preservation of STM will benefit other research areas. Furthermore, it may generate international prestige and, as such, provide an attractive venture for the Dutch Government to support.

> *'There is a question of principle to what extent the Dutch taxpayer should pay for the international role of KB. But this could be justified for three reasons: the overall costs of the e-Depot are still relatively small; most of the costs had already been made for preservation of Dutch imprint; and it is customary that the host nation of an international facility should pay a substantial amount of the cost.'*
>
> Cornelis van Bochove, Netherlands Ministry of Education, Culture and Science

KB should be aware that if this becomes a leading argument to support its international strategy and its preservation of international publications, it will impact the sources from which KB R&D and e-Depot receive their funding and the grounds on which they are evaluated. This funding is likely to be more competitive, and possibly more volatile, than that based on a legal obligation to store and preserve. Innovations of the kind that KB has produced (and may continue to produce) are likely to have a commercial value for which it could be encouraged to seek private funding. Eventually one could ask the question whether international preservation services and the extensive R&D required should remain the responsibility of a national library.

[53] For example, recent developments include the establishment of the European Alliance for Permanent Access of the Records of Science. Founding members include, *inter alia*: CERN, European Space Agency, European Science Foundation, Science and Technology Facilities Council (UK), Max Planck Gesellschaft (Germany), Centre National d'Etudes Spatiales (France), the British Library, Deutsche Nationalbibliothek and KB, the International Association of Scientific, Technical and Medical Publishers and several national coalitions for digital preservation. The Alliance aims to become a strategic partner for national governments and the European Commission to support the further development and implementation of their policies in the area of preservation and long-term access.

In addition to questions about KB's evolution from a national to an international remit, and from a library to a research centre, there are some concerns among stakeholders about KB's assumed duty to provide access to the public at large. Some feel that this should be expected from a public institution funded by taxpayers' money. In the traditional library, paper copies from its deposit collection have been accessible only on-site in order to avoid loss of the 'last resort copy'. The limitations of paper media have mitigated the risk of large-scale copyright breaches. With electronic copies this has changed, as have expectations about how to access them – i.e. from a distance over the Internet. Although open access is gaining ground and publishers are accommodating this development to some extent, they continue to indicate resistance to free, instant access to their most valuable assets and generally would oppose depositing content if their rights were not protected adequately.

Therefore, the question of public access may shift to the financing model, since the benefit to society of subsidising services to commercial publishers is questionable. Moreover, STM research itself is funded mostly by public means and the licensee of scientific journals (university library) is dependent on public subsidy. Thus if storage and preservation is also offered for free, publishers would benefit threefold (via research, university library subscriptions and preservation) from public monies. Publishers are aware of this, and some indicate that a natural allocation of responsibility would be for them to pay KB for providing perpetual access, whereas governments should address technological obsolescence.[54]

6.4 Need for a sustainable funding model

KB's main sponsor remains the Dutch Ministry of Education, Culture and Science. One argument for its supporting KB's expansion of its preservation activities into the international realm is that in practice it is difficult to distinguish national from international imprint publications. In addition, Section 4.4 showed that the average cost per additional object for upgrading to the international level is relatively small. If, on the other hand, the international aspect becomes a dominant cost factor for preservation activities, KB may need to look for additional sponsors to support its international efforts. Although clearly the European Commission does have an international (or at least European) outlook and would be willing to support centres of excellence in digital preservation and archiving, as well as the cooperation between such centres, its financial instruments are limited; it will not fund operational costs and uses open calls for proposals. Inevitably, such EC funding would be project-based and thus not sustainable for funding an e-Depot.

As stated in the previous section, there may be interest in other parts of the Dutch Government for supporting KB's pioneering role in this area. The prestige associated with being a world leader in digital preservation and an important player in the European context, as well as potential synergistic effects on other scientific disciplines, could entice

[54] However, it is questionable whether providing perpetual access and addressing technological obsolescence can be separated in this way.

the Ministry of Economic Affairs to support KB's efforts, alongside the Ministry of Education, Culture and Science.

KB is convinced that the synergy between R&D efforts and ongoing e-Depot operations has been key to its success; research at KB is linked directly with the development and maintenance of techniques, expertise and good practice in digital archiving and preservation. This notion is emphasised by the continuous need to monitor and adjust to new developments. To sustain this success, substantial R&D funds will continue to be required. A funding opportunity for KB would be to better capture the value that its services create. Although the market for preservation still seems rather small and the margins limited, clearly there are opportunities to charge fees. For publishers, perpetual access is a key element of their offering to their licensees. Effective preservation is a de facto insurance policy on STM content to which licensees have acquired the right of perpetual access. Without such a guarantee, publishers would not be able to sell their licenses, so this must be valuable to them. In parallel to delineating the market value of preservation, KB should provide a detailed overview of the costs of its e-Depot operations.

Hitherto there has been little evidence of a positive correlation between the technological robustness of preservation services, on the one hand, and the demand for such services, on the other. This is likely to be due to users' limited awareness and knowledge of these services. Several of the university libraries interviewed seem to think that Portico and CLOCKSS are sufficient alternatives. Moreover (as mentioned above), there is some distrust of national libraries. Some publishers fear national libraries' approach to intellectual property, whereas university libraries tend to see them as competitors. Both express doubts about national libraries' capabilities to deliver effectively in the case of trigger events.

There is already general acceptance among publishers that paying a fee for KB's archiving and preservation services would be justified. In the case of a trigger event, all parties agree that the insurance service provided by KB and the added costs that this would generate should be covered by the publishers. KB must analyse what the current value of this service is. It seems that willingness to pay is still limited; although several publishers expressed the need for some redundancy, different storage and preservation service providers are currently seen as interchangeable (and, in fact, most publishers have contracts with a number of such providers).

> *'If we decided to work with a small number of archives, then we would want to work with those which provided archiving services to the largest number of libraries. In spite of its excellent and pioneering work in this field, KB may not be regarded as a primary provider of archiving services by many libraries outside The Netherlands.'*
>
> Steven Hall, Wiley-Blackwell

For KB to increase its value, it needs to position itself better with university libraries in order to strengthen their demand. Also, it should endeavour to help improve general understanding of the issues and risks that surround preservation, or the lack of it. Thus KB has a branding challenge, to ensure that its name is associated with the best possible assurance of perpetual access, that it can actually deliver the services when they are required, and that it has no interest in exploiting the content with which it is entrusted.

In addition, the value of preserved collections is likely to grow following the long-tail theory, which will make preservation services more valuable to publishers – in which case, they may decide to develop their own preservation services or be willing to pay more for KB's services. This depends on the expected margins and technical barriers to entry into the preservation business. Such increase in value is likely to affect the publisher's concern with KB's duty to provide access and its ability to handle trigger events.

Many interviewees feel that a national library should not act beyond its traditional remit and that commercial exploitation of content – for example, by supplying 'pay-per-view' or access control – is undesirable, as it would undermine the relationship with the publishers, which remain the owner of the content.

Finally, given the trends described above, it is expected that the composition of KB's revenues (including public subsidies) may become more diversified. A larger share will depend on the services that it provides and the actual value that KB creates for its clients and society at large. For the full-scale launch of any preservation strategy and/or change in business model, KB would need to assess its cost base and analyse the market value of its services. We have attempted to gain insight into the costs of preservation and e-Depot in particular (see Section 4.4), but precise estimates are difficult to acquire. The lack of detailed financial implications of the supply and demand side of preservation complicates any process of strategic decision-making, where projections are required for the net present value of investments against future cash flows or other expected (societal) benefits that e-Depot may generate.

6.5 More transparent access rights policy

As previously mentioned, providing access to its collections is a core function of a (national) library. However, in the digital environment, this conflicts with the interests of the publishing industry, which retains copyright on the published content. There remain substantial disagreements as to when trigger events may occur and what their consequences are. Publishers are particularly concerned with preservation libraries' handling of trigger events. There are three models for this:

- access for licensees only;
- temporarily free to the world; or
- inherit on a subscribers list (such as Portico).

Not all publishers are comfortable with preservation models that grant open access (before copyright has expired) in case of a trigger event, as there is a risk of piracy and subsequent loss of control over content.

Equally, some observers indicate that research and university libraries may not like the idea that everyone has access to something for which they have paid. The ideal situation would be for a safe place to have access mechanisms in place for only those who have – or once had – license rights for digital content. Current solutions do not meet those needs. This would require establishing a register of access rights, licensees and former licensees, which is a highly cumbersome task.

> *'I would prefer a solution which, in the case of a trigger event, granted access only to those institutions which had licensed the content in the past, although I understand that this would be difficult. If the content is opened to the whole world without restriction, there is a real risk of piracy and once the trigger event is over, control of the content will be lost.'*
>
> Steven Hall, Wiley-Blackwell

To address at least some of this concern, KB should clarify its definition and conditions of trigger events and clarify what services can be expected under which circumstances. This should be communicated with libraries and publishers in order to allow them to anticipate the outcome of such events. The development of permanent services after trigger events potentially could be a cost-recovering activity for KB.[55] One such service could be for KB to act as the deposit library for migrated journals: KB could provide the back files to, or on behalf of, the archiving publisher or their successor. In addition, this service would be of interest to learned societies.

For KB, the issue of access to preserved collections may extend in future beyond the specific issues raised through the occurrence of trigger events.[56] Researchers and scholars are increasingly likely to expect the resources they need to be available and accessible online. Beyond their expectation for online access to individual institutions' holdings, scholars are likely to want increasing access to virtual, unbounded corpora that combine the holdings of multiple institutions. It is recognised that, for the most part, scholars are uninterested in where resources reside, except to the extent that credible stewardship helps to confer authenticity on those resources. Future scholarship is likely to depend increasingly on broadly-federated, virtualised collections and cross-corpus data mining.

Additionally, KB should evaluate the e-Depot access regime prior to a trigger event. Currently, e-Depot can be considered as a 'dim archive', as it is neither 'dark' nor 'light', only accessible to users on-site at KB's premises. While this policy is considered by KB to be non-financial compensation for its free service to publishers, it is met with scepticism by several stakeholder groups. On-site-only access to digital information on library premises is quickly becoming anathema and as such has barely any value. Furthermore, this ambiguous compensation scheme harms the transparency of the cost structure of preservation services. The compensation structure for archiving services should be set by the market value of such services and the costs related to managing access in case of trigger events. Access policies to such content would need to be addressed as a separate issue, and preferably should be based on regular licensing agreements with publishers.

[55] Costs of temporary access following a trigger event are to be covered by the publisher, see Section 4.3.2.

[56] To a certain extent, this is already the case in the DARE agreements between KB and Dutch universities to preserve the information published through their institutional repositories.

6.6 Coordination between initiatives and peers

The most prominent initiatives related to the preservation of e-journals that are currently in operation include KB's e-Depot, Portico and (C)LOCKSS. As discussed in Chapter 4, these efforts have distinct mandates, funding sources, business models, temporal outlooks, preservation strategies, arrangements with e-journal publishers and relationships with scholars. Most observers appear to think that it would be to the mutual benefit of these initiatives to establish relationships with each other. However, it is unclear just what these relationships should be.

Since both Portico and CLOCKSS are so new, KB is the only one of these three efforts that has a track record and consequent credibility as a reliable preservation institution. As a national library, KB is seen as having a longer-term perspective and no financial motives, as well as an orientation toward scholarly access. Notwithstanding the scepticism of some stakeholders toward KB (as discussed in the previous section), these perceived advantages earn it a high degree of credibility.

Therefore, it is important that any relationship that KB forges with Portico and/or (C)LOCKSS avoids compromising any of this credibility – and indeed, should enhance it, if at all possible. This suggests that KB should guard against diluting any of its key advantages in such a relationship, i.e. its financial independence from publishers, multi-pronged, long-term preservation perspective and orientation toward scholarly access. Some of these advantages may be affected by other factors beyond KB's control – in particular, as noted above, KB may be forced to compromise its provision of access to scholars due to publishers' increased interest in long-tail revenue streams coupled with extended copyright protection – but KB should be careful to retain or augment its advantages to the greatest possible extent.

These considerations imply that the most valuable relationships that KB might forge with Portico and/or (C)LOCKSS may be those that provide cross-monitoring and that (partially) duplicate the digital collection. Mutual agreements that would cover these aspects should enhance both parties' credibility. However, KB should be wary of compromising its own access rights or other capabilities in forming such agreements. Beyond this, it would be mutually beneficial to arrange for the free and open sharing of useful concepts such as the development of preservation procedures, the administration of 'dark' or 'dim' archives, insights, techniques and experiences related to the technology of preservation, or Fair Use access restrictions that would be acceptable to both publishers and libraries. All parties are likely to benefit from such agreements. However, KB should recognise the uniqueness of its current position in the preservation arena, and be sure that any agreements it undertakes actually preserve or enhance its position.

KB is actively pursuing a strategy of cooperation with its fellow national libraries, such as the British Library and the Deutsche Nationalbibliothek, to develop a network of safe places as a pre-competitive alliance which would provide mutual benefits. One university representative from the UK expressed the view that only two or three bodies in Europe – including KB – would be in a position to be part of such a global network of about five to seven safe places. Most stakeholders embrace this vision and see such a network as a way to guarantee preservation with a healthy level of redundancy, while at the same time avoiding excessive and costly overlap. Also, it would be natural for the institutions comprising a safe

places network to share costs and technology, exchange best practice and cooperate in agreeing on common standards. Publishers are in favour of this, but they deplore the fact that there is no leadership in this development and no effective platform for discussion between the main (national) preservation libraries. This is illustrated in the following quote:

> *'KB's idea of 'trusted digital repositories' will need further thought. The model proposed by OCLC and RLG is too complicated. None of us is ready to decide yet.'*
>
> Richard Boulderstone, British Library

6.7 Specify scope of preservation

As stated in the introduction, the real disruptions in the next decade(s) are unlikely to come from the current players. As long as scientific publications are at the heart of the business model, publishers, university libraries and archives will evolve together following technological and market trends. Thus the strategies and approaches described above are well suited to deal with the changes in this medium-term context. The major disruptions in the long term are likely to come from fundamental changes in scholarly communication – how it takes place, what it produces, how it is embedded in tenure and reward systems, etc., in combination with new ways of publishing such as open access.

Commercially-exploitable journal publications are still the main unit of scholarly communication. However, this is likely to change in the next decade. Slowly but surely, the use of non-page-image objects is intensifying, including:

- dynamic webpages;
- animations;
- video and other multimedia;
- databases;
- geographic information systems;
- models and simulations;
- finer-grained units of information and embedded objects;
- virtual compound objects and inherently digital objects, including script-generated webpages and executable models;
- visualisations; and
- programs of all kinds.

Although these new types of objects do not constitute a dramatic percentage of the scholarly record yet, they seem likely to become increasingly numerous and important over the next 10 to 20 years, if not well before then.

KB as a leader in digital preservation is at the forefront of addressing these disruptive developments. A major distinction between KB and most other members of the preservation community may be the fact that KB's approach is founded on solid, flexible

and experimentally-informed technical grounds. In order to maintain this advantage, KB will need to reassess its technical assumptions continuously in the light of continuing developments in scholarly communication, and it will need to continue to conduct successively more sophisticated and larger-scale experiments to verify the adequacy of its strategy. This may require accelerating its development of preservation techniques (such as emulation) that can cope with inherently digital formats.[57]

> *'While some dissemination functions will stay, new channels for discussing and sharing knowledge and information will become citable and thus part of the collective research output. These channels should be included in a national library's archive of scholarly publications.'*
>
> Reinhard Altenhöner, Deutsche Nationalbibliothek

Change is not limited to traditional (commercial) scientific publication (see Section 5.2.2). Shifts in scholarly communication toward informal publication, open-access publication, institutional publication, discipline-based publication and self-publication also will affect KB's relationships with traditional publishers over time. In addition, KB will have to establish new relationships with universities, new discipline-based publishers, self-publication services, web hosting and other Internet services and possibly even information hubs (such as Google), as well as individual authors.

KB will be challenged – first and foremost in its public service role as archive for national STM and other published digital output – to identify which elements of the scholarly record it should target for preservation and which ones it should not, especially since the preservation of this material will not be funded through a market mechanism. In fact, the selection of material worthy of archiving could become more time-consuming than archiving everything.

The safe places model offers a suitable solution for capturing such a wide variety of content from such a wide range of sources for archiving. Alternatively, institutional archives could be preserved through a local emulation or migration solution. Access could be facilitated then through harvesting and search engines. The library community could take responsibility for assuming (part of) these costs, for example for the journals listed in the Directorate of Open Access Journals.

Publication from smaller publishers and less prestigious journals would benefit from the safe places model, since their preservation is not yet guaranteed. For KB (and its peers) it is important to consider how to encourage small publishers to participate in archiving schemes. An option would be to make depositing material as easy as possible, for example by establishing a universal deposit system supported by the network of safe places. At the same time, effective guarantees must be in place to protect publishers' copyright, thus creating an environment of trust. KB could provide guidelines for doing this, facilitate partnerships, etc. The Scholarly Publishing and Academic Resources Coalition (SPARC) indicates a willingness to collaborate on this.

[57] There are already ongoing R&D developments at KB in this direction.

Obviously, the choice to include other kinds of content in addition to STM publications represents the much broader societal question of what needs to be kept for future use: what is the future of archiving in a time where information is being generated everywhere and at unprecedented speed? New generations may have different priorities for preservation. Data mining and search technology may suffice for many purposes in the future, and should capture the current, ongoing and recent production of a great variety of scientific output and scholarly communication. However, this does not (yet) guarantee preservation and perpetual access. Thus, the question remains as to whether it would be desirable to include other content beyond traditional STM publications in preservation services, and if so, how this can be achieved and what KB's role would be. KB should continue its active pursuit of such possibilities by means of experiments involving the preservation of web content and other non-traditional forms of publication.

6.8 Robust planning of an e-Depot strategy

For KB to take the next steps in planning and rolling-out its e-Depot strategy, it needs to understand its cost base and the value of the services that it provides. Also, it must factor in endogenous developments in the marketplace for preservation and scholarly communication, as well as advances in technology. Finally, it should regularly monitor and evaluate its progress against its strategic objectives.

To address the exogenous factors that will affect the future of KB's strategy in a fast-changing global setting, KB should consider testing its strategic options against a set of scenarios. Chapter 5 elaborated four scenarios for the future of digital archiving and preservation. These are not predictions, but end-of-spectrum examples that are useful for illustrating the range of plausible futures in the next 10 to 20 years. They provide a framework to consider the uncertainties, key drivers and policy levers that will determine the context in which KB has to operate in the near, mid- and long-term future.

Scenarios are useful because they raise awareness and shed new light on current strategic debates. More important in the present context is that by using multiple scenarios, it becomes possible to test policy recommendations for their robustness. If an option appears to be effective in several, highly different scenarios, this implies that the option is robust. For options that are not robust, it is equally significant to understand under which circumstances they are not effective. Thus, the identified scenarios allow KB to assess its strategy on various factors that play out in different ways, in different markets and against different time horizons. Scenarios also help the organisation to be sensitive to future trends and allow effective responses once these trends begin to emerge, since some of the conceptual groundwork has taken place already. Furthermore, scenarios may be used to communicate a vision to KB's stakeholder community and to rally stakeholders around common approaches to likely future challenges and uncertainties.

In order to support robust planning, KB could facilitate an assessment of the need for preservation in each scenario on a variety of aspects: organisational, financial, institutional, technical, etc. Using a technique called scenario gaming, scenario workshops are organised with internal staff, key experts and stakeholders to support this assessment. This would enable KB and the wider community to delineate common elements that are robust in all

scenarios: the strategy should be built on these robust elements. Subsequently, the elements that are unique for each strategy will be identified. These enable the organisation to define decision milestones for identifying exogenous developments and their likely impacts, creating 'pathways to the future'.

In Chapter 5 we identified two dimensions, along which we have positioned the example scenarios for the future of preservation: the evolving nature of scholarly dissemination and the nature of demand for digital preservation. Besides these exogenous factors (which are assumed to be beyond KB's control), there are endogenous factors (which KB can influence). Additionally, decisions on the geographic scope (national versus international research output) or content scope (specific, STM versus general, all digital content) affect KB's strategic options in the national and global marketplace for digital preservation.[58] Combining the endogenous factors with the example scenarios in Chapter 5 may lead to various plausible strategic avenues for KB.

Once the strategy is set and the milestones are defined, it is important to monitor progress in the execution of the strategy and to reassess continuously the extent to which exogenous factors are playing out in the marketplace. Thus indicators should be developed to tie KB's performance to its strategic objectives. Milestones allow the organisation to determine what scenario is actually unfolding and what unforeseen elements must be taken into account in going forward. This would ensure that the strategy itself remains adjusted to the changing context in which KB operates.

6.9 Conclusion

The future outlook for digital archiving and preservation is uncertain, and there is no accepted single solution to the challenge of long-term preservation of scholarly output. KB has been among the first to address this challenge through the development and implementation of its e-Depot. We feel that KB's strategic choices involve three critical assumptions about the uncertain future of digital preservation. Without implying that KB is unaware of these assumptions or is not taking steps to address them, we nevertheless feel that they deserve to be made explicit:

1. For its international e-Depot, KB aims to sign archiving agreements with at least the 20 to 25 largest international publishing companies which produce almost 90 percent of the world's electronic STM literature. The assumption underlying this objective is that large traditional publishers will continue to be the main providers of this scholarly content, although KB is also pursuing other sources of STM literature, including institutional publication, self-publication and web publication.
2. KB believes that it is essential to preserve digital objects in their authentic form, including not only their content but also the full behaviour of their formats, since it is impossible to predict which attributes of an object may be important in the future. Consequently, KB considers preservation of the original object (or something very

[58] Again, these are not discrete options, but part of a continuum of strategic developments.

close to the original) to be essential, along with the ability to preserve the original behaviour of the object.

3. KB assumes that government funding of its tasks, including digital preservation, will continue into the long-term future.

These and other assumptions underlying KB's strategy hold relatively well for the short and medium term and are broadly supported by the main stakeholders. KB is in a good position to participate in an international network of safe places. For such a network to be implemented, leadership is required, which KB may be able to provide, given its position as a not-for-profit public institution from a small country with a reputation in digital preservation, relevant expertise and ongoing R&D efforts.

6.9.1 The short term

Despite the general endorsement of its strategy, there remain a number of issues which KB should consider in the short term. Again, we do not imply that KB is unaware of these issues or is not already taking steps to address them:

- Continue efforts to build trust and awareness in the wider international stakeholder community – especially among university libraries – of the importance of long-term digital preservation, the nature of KB's unique approach and its abilities to deliver the expected services over time. In doing this, KB should target organisations that have yet to devise a strategy for digital preservation, including university libraries in North America and Asia;
- Continue to assess its role as a national library and its core function of serving as a trusted steward for an authentic scholarly record, which requires independence of commercial interests, and balance this with other activities that are within the scope of the international strategy of e-Depot, such as preserving international imprint, performing consultancy and serving as a centre of research excellence;.
- Review its current funding model to achieve a more diversified and sustainable financial basis that is more attuned to the full range of e-Depot's activities, while accepting that public funding will continue to be indispensable for the core of KB's activities and while guarding against diluting any of its key advantages, i.e. its financial independence from publishers, its multi-pronged, long-term preservation perspective and its orientation toward scholarly access.. To facilitate this review, establish the market value of its services, and develop a detailed overview of the costs of its e-Depot operations;.
- Provide full transparency and communicate its policy, conditions and definitions of trigger events. Ensure that all access rights, whether on-site or following trigger events, are dealt with and priced based on the value of the services provided and their cost, rather than treating access rights as compensation in kind for preservation services to publishers.;

6.9.2 The medium term

For the medium term, KB should monitor external developments and adjust its approach as necessary:

- Continuously review emerging technological developments and accelerate development of preservation techniques that can cope with inherently digital formats;
- Monitor developments in scholarly dissemination and engage in the broader societal debate about the boundaries of the scholarly record;
- Assess which components of scholarly output should be selected for preservation into perpetuity and which should not.

6.9.3 The long term

Over the long term, the assumptions underlying KB's strategy are not necessarily robust in all plausible scenarios for the future. The future for digital preservation is highly uncertain due to developments in: scholarly dissemination and publishing; fragmentation of dissemination through a range of formal and informal channels versus further consolidations of traditional publishers; and the priority among stakeholders and funding agencies of authentically archiving and preserving inherently digital artefacts. To address these uncertainties, KB should continuously reconsider its assumptions and take into account possible and even unlikely future developments:

- Initiate a process of robust strategic planning through testing its strategic assumptions against a set of scenarios, in order to help address the uncertainties in the market for long-term digital preservation of scholarly output;
- Interact with internal and external stakeholders and actively engage them in KB's process, in order to address uncertainties and develop a shared vision on how to approach these uncertainties

REFERENCES

References

Albert, Karen (2006) Open Access: Implications for Scholarly Publishing and Medical Libraries. *Journal of the Medical Library Association* 94(3): 253–62.

Balistier, Thomas (2000) *The Phaistos Disc: An Account of its Unsolved Mystery*. Mahringen: Verlag.

Beagrie, Neil (2003) *National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands and the United Kingdom and of Related International Activity*. Washington, DC: Council on Library and Information Resources and Library of Congress. Available at: http://www.clir.org/PUBS/reports/pub116/pub116.PDF

Boyce, Peter (1996) A Successful Electronic Scholarly Journal from a Small Society. *UNESCO Expert Conference on Electronic Publishing in Science and its WG2 Workshop*. Paris: ICSU Press. Available at: http://www.aas.org/~pboyce/epubs/icsu-art.html

Butler, Declan (2006) Open-access Journal Hits Rocky Times. *Nature* 441: 914.

Dewatripont, Mathias, Ginsburgh, Victor, Legros, Patrick, Walckiers, Alexis, Devroey, Jean-Pierre, Dujardin, Marianne, Vandooren, Françoise, Dubois, Pierre, Foncel, Jerome, Ivaldi, Marc and Heusse, Marie-Dominique (2007) *Study on the Economic and Technical Evolution of the Scientific Publication Markets in Europe.* Commissioned by DG-Research, European Commission. Available at: http://ec.europa.eu/research/science-society/PDF/scientific-publication-study_en.PDF

Deweerd, Harvey (1973) A Conceptual Approach to Scenario Construction, P-5087. Santa Monica, CA: RAND Corporation.

Duhoux, Yves (1977) *Le disque de phaestos*. Leuven: Edition Peeters.

Dutch National Archives (2005) *Costs of Digital Preservation.* Available at: http://www.digitaleduurzaamheid.nl/bibliotheek/CoDPv1.PDF

Electronic Publishing Services (2006) *UK Scholarly Journals: 2006 Baseline Report. An Evidence-based Analysis of Data Concerning Scholarly Journal Publishing*: Final report. Available at: http://www.rin.ac.uk/data-scholarly-journals

Elsevier (2003) Elsevier Launches *The Lancet* Backfiles on ScienceDirect. Press release. Available at: http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_00034

European Commission (2007) *Staff Working Paper Accompanying the Commission Communication on Scientific Information in the Digital Age: Access, Dissemination and Preservation.* Available at: http://ec.europa.eu/research/science-society/document_library/PDF_06/comm-native-com-2007-0056-1-divers_en.PDF

Fenton, Eileen (2006) Preserving Electronic Scholarly Journals: Portico. *Ariadne* 47(April). Available at: http://www.ariadne.ac.uk/issue47/fenton/

Frazier, Kenneth (2001) The Librarians' Dilemma: Contemplating the Costs of the 'Big Deal'. *D-Lib Magazine* 7(3). Available at: http://www.dlib.org/dlib/march01/frazier/03frazier.html

Gooden, Paul, Owen, Matthew, Simon, Sarah and Singlehurs, Louise (2002) *Scientific Publishing: Knowledge Is Power.* Equity research, media. London: Morgan Stanley.

Guthrie, Kevin and Schonfeld, Roger (2004) What Do Faculty Think of Electronic Resources? Findings from the 2003 Academic Research Resources Study. Paper presented at the Coalition for Networked Information Taskforce Meeting, Alexandria, VA, 16 April.

Harley, Diane, Earl-Novell, Sarah, Arter, Jennifer, Lawrence, Shannon and Judson King, C. (2006) *The Influence of Academic Values on Scholarly Publication and Communication Practices,* Center for Studies in Higher Education, University of California–Berkeley. Available at: http://repositories.cdlib.org/cshe/CSHE-13-06

House of Commons, Science and Technology Committee (2004) Scientific Publications: Free for all? *Tenth Report of Session 2003-04,* Volume I: Report, HC 399-I. London: The Stationery Office.

Hunter, Karen (2007) The End of Print Journals: (In)Frequently Asked Questions. *Journal of Library Administration* 46(2):119–32.

ICTU (2003) *Van digitale vluchtigheid naar digitaal houvast Bewaren van spreadsheets.* The Hague, November. Available at: http://www.digitaleduurzaam heid.nl/bibliotheek/docs/Bewaren_van_spreadsheets.pdf

International Association of Scientific, Technical and Medical Publishers (STM) (2007) Brussels Declaration on STM Publishing. Available at: http://www.stm-assoc.org/brussels-declaration/

Joint Information Systems Committee (JISC) (2005) *Open Access.* Briefing paper. Available at: http://www.jisc.ac.uk/publications/publications/pub_openaccess.aspx

Joint Information Systems Committee (JISC) (2007) Conference Addresses: *Archiving and Preservation of e-Journals.* London, 28 March. Available at: http://www.jisc.ac.uk/news/stories/2007/03/news_e-journals.aspx

Kahn, Herman and Weiner, Anthony (1967) *The Year 2000: A Framework for Speculation on the Next Thirty Years*. London: Macmillan.

Kahn, Herman, Brown, William and Martel, Leon (1976) *The Next 200 Years: A Scenario for America and the World*. New York: William Morrow and Co.

Kenney, Anne, Entlich, Richard, Hirtle, Peter, McGovern, Nancy and Buckley, Ellie (2006) *e-Journal Archiving Metes and Bounds: A Survey of the Landscape*. Washington, DC: Council on Library and Information Resources.

Kobrak, Fred and Luey, Beth (eds) (2002) *The Structure of International Publishing in the 1990s*. New Brunswick, NJ: Transaction Publishers.

Koninklijke Bibliotheek (2002) *The Road to e-Depot at the Koninklijke Bibliotheek.* Available at: http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties_edepot.html

Koninklijke Bibliotheek (2007) *About KB – The e-Depot: An Introduction.* Available at: http://www.kb.nl

Library of Congress (2007) *Metadata Encoding & Transmission Standard.* The Library of Congress website. Available at: http://www.loc.gov/standards/mets/. Accessed on: 25 September 2007.

McLeod, Rory, Wheatley, Paul and Ayris, Paul (2006) *Life-cycle Information for e-Literature: Full Report from the LIFE Project*. London: LIFE Project. Available at: http://www.ucl.ac.uk/ls/life/1/documentation.shtml

Mabe, Michael (2003) The Growth and Number of Journals. *Serials* 16(2): 191–7.

Mabe, Michael (2006) Journal Futures: How Will Researchers Communicate as the Internet Matures? Paper presented at the *Fiesole Collection Development Retreat Series*, Lund, August. Available at: http://digital.casalini.it/retreat/2006_docs/mabe.PDF

Mark Ware Consulting (2006) *Scientific Publishing in Transition: An Overview of Current Developments*. White Paper commissioned by STM and ALPSP, 14 September. Available at: http://www.stm-assoc.org/helpful-articles-reports-messa/

Moed, Henk (2005) Statistical Relationships between Downloads and Citations at the Level of Individual Documents within a Single Journal. *Journal of the American Society for Information Science and Technology* 56(10): 1088–97.

Muir, Adrienne (2001) *Legal Deposit of Digital Publications: A Review of Research and Development Activity. First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01)*. New York: ACM Press.

National Institutes of Health (NIH) (2003) *NIH Public Access: Questions and Answers. Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research.* Available at: http://publicaccess.nih.gov/publicaccess_QandA.htm

Oltmans, Erik and Kol, Nanda (2005) A Comparison Between Migration and Emulation in Terms of Costs. *RLG DigiNews*, issue index, 15 April. Available at: http://www.rlg.org/en/page.php?Page_ID=20571#article0

Oltmans, Erik and Lemmen, Adriaan (2006) e-Depot, National Library of the Netherlands. *Serials* 19(1): 61–7.

Oltmans, Erik and van Wijngaarden, Hilde (2006) The KB e-Depot Digital Archiving Policy. *Library Hi Tech* 24(4): 604–13.

Organization for Economic Cooperation and Development (OECD) (2006) *OECD Science, Technology and Industry Outlook*. Paris: OECD. Available at: http://www.oecd.org/dataoecd/39/19/37685541.PDF

Organization for Economic Cooperation and Development (OECD) (2007) *Main Science and Technology Indicators, 2007*. Available at: http://www.oecd.org

Pearce-Moses, Richard (2005) *A Glossary of Archival and Records Terminology*. Chicago, IL: Society of American Archivists. Available at http://www.archivists.org/glossary/term_details.asp?DefinitionKey=231

Petersen, John (2000) *Out of The Blue: How to Anticipate Big Future Surprises*. Lanham, MD: Madison Books.

Powell, David (2004) Publishing Output to 2020. In: Electronic Publishing Services (ed.) *The Future of Print and Electronic Publishing Output Worldwide*, 29 January. Abstract available at http://www.bl.uk/about/articles/pdf/epsreport.pdf

RAND Europe (1997) *Scenarios for Examining Civil Aviation Infrastructure Options in The Netherlands.* Unrestricted draft, DRU-1513/VW/VROM/EZ.

Reding, Viviane (2007) Scientific Information in the Digital Age: How Accessible Should Publicly-funded Research Be? Speech by the Commissioner for Information Society and Media, *Conference on Scientific Publishing in the European Research Area Access, Dissemination and Preservation in the Digital Age,* Brussels, 16 February.

Research Libraries Group (2002) *Trusted Digital Repositories: Attributes and Responsibilities*. RLG-OCLC Report. Available at: http://www.rlg.org/longterm/repositories.PDF

Research Libraries Group (2007a) *Trustworthy Repositories. Audit & Certification: Criteria and Checklist.* February. Available at: http://bibpurl.oclc.org/web/16712

Research Libraries Group (2007b) *Criteria for Measuring Trustworthiness of Digital Repositories and Archives: an Audit & Certification Checklist*, draft version 1, RLG-NARA taskforce, January.

Rosenthal, David S. H., Lipkis, Thomas, Robertson, Thomas S. and Morabito, S. (2005) Transparent Format Migration of Preserved Web Content, *D-Lib Magazine* 11(1). Available at: http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

Rothenberg, Jeff (1995) Ensuring the Longevity of Digital Documents. *Scientific American* 272(1): 42–7.

Rothenberg, Jeff (1999) *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. A Report to the Council on Library and Information Resources, January. Washington, DC: Council on Library and

Information Resources. Available at: http://www.clir.org/pubs/reports/rothenberg/pub77.PDF

Rothenberg, Jeff (2000a) *Using Emulation to Preserve Digital Documents*. The Hague: Koninklijke Bibliotheek. Available at: http://www.konbib.nl/kb/pr/fonds/emulation/emulation-en.html

Rothenberg, Jeff (2000b) Preserving Authentic Digital Information. In: Council on Library and Information Resources (ed.) *Authenticity in a Digital Environment*. Washington, DC: Council on Library and Information Resources, pp. 51–68. Available at: http://www.clir.org/pubs/reports/pub92/pub92.PDF

Rothenberg, Jeff (2006) Renewing The Erl King. *Millennium Film Journal: Hybrids* 45–46(Fall): 20–51. Available at: http://www.bampfa.berkeley.edu/about/ErlKing Report.PDF

Rothenberg, Jeff and Bikson, Tora (1999) *Digital Preservation: Carrying Authentic, Understandable and Usable Digital Records Through Time*. Report to the Dutch National Archives and Ministry of the Interior. The Hague: RAND Europe. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.PDF

Rothenberg, Jeff and Slats, Jacqueline (2003) *Emulation: Context and Current Status*. Dutch Archives Digital Preservation Testbed and RAND Europe. Available at: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.PDF

Rowlands, Ian and Nicholas, Dave (2005) *New Journal Publishing Models*. Report. London: Centre for Information Behaviour and the Evaluation of Research (CIBER).

Rowlands, Ian, Nicholas, Dave and Huntingdon, Paul (2004) *Scholarly Communication in the Digital Environment: What Do Authors Want?* Report. London: Centre for Information Behaviour and the Evaluation of Research (CIBER).

Schwartz, Peter (1991) *The Art of the Long View: Planning for the Future in an Uncertain World*. New York: Doubleday.

Simba Information (2007) Global STM Publishing 2007–2008. Executive Summary. Available at: http://www.simbanet.com/publications/summaries/PDFs/Exec Summary_GSTM2007.PDF

Smith, Robert (1964) *Some Thoughts on Scenarios*, D-12250-PR. Santa Monica, CA: RAND Corporation.

Spinellis, Diomidis (2002) The Decay and Failures of Web References. *Communications of the ACM* 46(1): 71–7.

Stanescu, Andreas (2005) *Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology*. Digital Collections and Preservation Services, OCLC Systems and Services. *International Digital Library Perspectives* 21(1): 61–81.

Steenbakkers, Johan (2003) Permanent Archiving of Electronic Publications. *Serials* 16(1): 33–6.

Tenopir, Carol (2003) *Use and Users of Electronic Library Resources: An Overview and Analysis of Recent Research Studies*. Washington, DC: Council on Library and Information Resources. Available at http://www.clir.org/pubs/abstract/pub120abst.html

Tenopir, Carol and King, Donald W. (2000) *Towards Electronic Journals: Realities for Scientists, Librarians and Publishers*. Final report, January. Washington, DC: Special Libraries Association.

The Wellcome Trust (2006) Position Statement in Support of Open and Unrestricted Access to Published Research. Available at: http://www.wellcome.ac.uk/doc_WTD002766.html

Thomson Scientific (2007) *ISI Thomson Journal Citation Report*s. Available, via ISI Web of Knowledge, at: http://isiwebofknowledge.com/

Trauth, Michael (1990) The Phaistos Disc and the Devil's Advocate. On the Aporias of an Ancient Topic of Research. *Glottometrika* 12: 151–73.

van der Werf, Titia (2000) *The Deposit System for Electronic Publications. A Process Model.* NEDLIB Report Series no. 6. The Hague: Koninklijke Bibliotheek.

van Drimmelen, Wim (2004) Universal Access through Time: Archiving Strategies for Digital Publications. *Libri: International Journal of Libraries and Information Services* 54(2). Available at: http://www.librijournal.org/2004-2toc.html

van Notten, Phillip, Sleegers, Am and van Asselt, Marjolein (2005) The Future Shocks: On Discontinuity and Scenario Development. *Technological Forecasting and Social Change* 72(2): 175–94.

Versita (2006) STM *Publishing Industry and Market Overview. Central European Science Journals.* Available at: http://versita.com/UserFiles/File/STM%20Publishing%20Industry%20and%20Market.PDF

Vinocur, John (1999) Strauss-Kahn's Resignation Is a Major Blow for Socialists: Finance Chief Quits In French Scandal. *International Herald Tribune*, 3 November. Available at: http://www.iht.com/articles/1999/11/03/france.2.t_0.php

Waller, Andrew and Bird, Gwen (2006) 'We Own it': Dealing with 'Perpetual Access' in Big Deals. *Serials Librarian* 50(1–2): 179–96.

World Wide Web Consortium (2001) *Semantic Web*. Available at: http://www.w3.org/2001/sw/

APPENDICES

Appendix A: Summary of key assumptions underpinning the e-Depot strategy

The following assumptions were extracted and interpreted from KB's international e-Depot strategy and consequently endorsed by KB.

- The origin and location of networked digital information are losing their relevance in archiving. As a consequence, eventually the regulatory obligations of national libraries, which govern their national archiving function, will have to change.
- The market for the archiving and preservation of STM data is underdeveloped. Major investments are needed to develop and maintain the necessary infrastructures.
- The revenues which can be derived from providing e-Depot services are quite limited compared to the costs of operating an e-Depot and the investments needed to maintain and build the underlying technologies. Thus the economics are not attractive enough to permit an open market.
- Because of the investment costs involved and the unattractive economics, the development and operation of e-Depots will depend on government or private subsidies into the foreseeable future.
- The high costs of investment and operations create substantial economies of scale and barriers to entry.
- Publishers want to spread risks but also are subject to geopolitical considerations.
- For their e-Depot needs, publishers want a network consisting of a small number of trusted repositories. For safety reasons, publishers will entrust the same content to multiple archives. In this context, 'trusted' is defined by a set of generally-accepted criteria (see Research Libraries Group 2007a, 2007b).
- The structure of relations between publishers, users and financiers will create an environment which favours cooperation between e-archival providers and limits competitive behaviour.
- Acceptability standards, economics and the need for security will favour the emergence of a small group of e-Depot suppliers operating in an oligopolistic environment. However, these will have to meet a set of generally-accepted criteria for trusted repositories.

- KB's e-Depot ranks among the leaders in the development, introduction and operation of digital repositories.
- KB's leading position in digital archiving and preservation is in the national interest of The Netherlands.
- KB's ambitions in the digital world have a strong positive impact on the skills and motivation of KB staff and thus benefit all aspects of its operations.

Appendix B: List of interviewees

Interviewee	Institution	Function	Country
Donald J. Waters	Andrew Mellon Foundation	Program Officer, Scholarly Communication	USA
Matthew Cockerill	BioMed Central	Publisher	UK
Steven Hall	Blackwell Publishing	Journals Sales and Marketing Director	UK
Richard Boulderstone	The British Library	Director of e-Strategy and Information Systems	UK
Ronald Milne	The British Library	Director of Collections at the British Library	UK
Andy Williams	Cambridge University Press	Project leader, Archiving Strategy for the Next 10 Years	UK
Amy Friedlander	Council on Library and Information Resources	Director of Programs	USA
Abby Smith		Independent consultant	USA
Maria Heijne	Delft University of Technology	Librarian	The Netherlands
Karen Hunter	Elsevier Science	Vice-President	USA
Nick Fowler	Elsevier Science	Director of Strategy	The Netherlands
Carlos Morais Pires	European Commission, Directorate-General Information Society and Media	Head of Sector 'Scientific Data Infrastructures' Unit F3 - GEANT and e-Infrastructures	EU
Wim Jansen	European Commission, Directorate-General Information Society and Media	Scientific Officer, GÉANT and Users' Communities	EU
Dale Flecker	Harvard University	Associate Director for Planning and Research	USA
Raymond van Diessen	IBM Business Consulting Services	Managing Consultant	The Netherlands

Interviewee	Institution	Function	Country
Reinier A.M. van Langen	IBM Global Business Services	Partner, Public Sector	The Netherlands
Joost Geise	IBM Global Education Industry Solution Unit		The Netherlands
Ute Schwens	Deutsche Nationalbibliothek	Director	Germany
Reinhard Altenhöner	Deutsche Nationalbibliothek	Head, IT Department	Germany
Frank L.A. van Oudenhove	IBM Nederland	Delivery Project Executive, AMS Netherlands	The Netherlands
Herman P. Spruijt	International Publishers Association	Vice-chairman	Switzerland
Kathleen Keane	Johns Hopkins University Press	Director	USA
Hilde van Wijngaarden	Koninklijke Bibliotheek	Head, Digital Preservation	The Netherlands
Erik Oltmans	Koninklijke Bibliotheek	Head, e-Depot	The Netherlands
Hans J. Jansen	Koninklijke Bibliotheek	Director, Research and Development	The Netherlands
Johan F. Steenbakkers	Koninklijke Bibliotheek	Director, e-Strategy and Property Management	The Netherlands
Wim van Drimmelen	Koninklijke Bibliotheek	General Director	The Netherlands
Frans H.J. Visscher	Korn/Ferry (formerly Reed-Elsevier)	Formerly Chief Operating Officer, Reed	The Netherlands
Paul Ayris	LIBER	Director of Library Services and Copyright Officer, UCL, and Head of LIBER Access Division	UK
Deanna Marcum	Library of Congress	Associate Librarian for Library Services	USA
Martha Anderson	Library of Congress	Digital Projects Manager	USA
Joachim Driessen	Malmberg	General Manager (formerly Head of Strategy of Wegener; Head of Legal Publishing at Kluwer)	The Netherlands
Cornelis van Bochove	Ministry of Education, Culture and Science	Director, Department for Research and Science Policy	The Netherlands
Herman J. Bruggink	Nyenrode Business University	President	The Netherlands
Alice Keller	Oxford University Library	Head of Collection Management	UK
Eileen Fenton	Portico	Executive Director	USA

Interviewee	Institution	Function	Country
Mary Sauer-Games	ProQuest CSA	Vice-president of Publishing	USA
Dan Burnstone	ProQuest CSA	Publishing Director	UK
Robin Dale	Research Libraries Group	Co-chair, Taskforce on Digital Repository Certification	USA
Miao Qihao	Shanghai Library	Deputy Director	China
Xu Qiang	Shanghai Library	Director of System and Network Centre	China
David Prosser	SPARC Europe	Head of Systems Strategy	UK
Peter Hendriks	Springer	President Global Sales and Marketing	The Netherlands
David Rosenthal	Stanford University	Chief Scientist, LOCKSS Program	USA
Victoria A. Reich	Stanford University	Director, LOCKSS Program	USA
Pieter Bolman	International Association of Scientific, Technical and Medical Publishers	Adviser and former Chief Executive Officer	UK
Kurt de Belder	University of Leiden	Librarian	The Netherlands
Paul Courant	University of Michigan	Professor of Public Policy	USA
Dave Thomson	The Wellcome Trust	Web Archiving Project Officer	UK
Robert Kiley	The Wellcome Trust	Head of Systems Strategy	UK

Appendix C: The technological basis for digital preservation

Although many factors affect the outcome of the preservation enterprise, having a sound technological basis for digital preservation is fundamental. Without such a basis, all the rest will be wasted effort, since the digital information in question will become inaccessible or uninterpretable (Rothenberg 1999).

Preserving the bitstream

The tip of the iceberg of this problem is the preservation of the bitstream that comprises every digital object. Such bitstreams must be copied onto new storage media whenever older media wear out or become obsolete. This 'media migration' is now a well-understood and well-accepted aspect of digital preservation, and although it has significant cost implications, carrying it out is reasonably straightforward. In order to be safe, multiple copies of each bitstream ideally should be placed in multiple geographically-separated, institutionally-independent repositories in order to guard against natural, economic, criminal, ideological and political threats. Periodic cross-auditing of these replicated repositories should ensure that their contents remain valid and uncorrupted.[59]

Rendering for human interpretation

However, the core of the preservation problem lies in the fact that a bitstream is unintelligible without further interpretation. That is, every digital object is encoded and must be interpreted in order to render its content into a form that humans can perceive and use (for convenience, we will refer to this as being 'human-readable', even though digital objects may be viewed, heard and experienced in many other ways in addition to being read). Although in principle this rendering process can be carried out by a human, in practice, this is infeasible due to the size and complexity of the interpretation task for all but the tiniest and most trivial of digital objects. In order to render a digital object human-readable, it is necessary to run software (i.e. a computer program) on a suitable computer in order to interpret the object's bitstream.

The encoding of each digital object obeys a particular format, which specifies a set of rules for interpreting its bitstream. Documents in a given format can be created and rendered by at least one computer program and often can be rendered by several or even many other programs. However, due to the rapid evolution of computer science, a given format

[59] Kenney *et al* (2006: 51) use the expression 'Trust, but verify' to refer to such cross-validation.

becomes obsolete in a relatively short time. The lifetime of a given format is typically only a few years, since every second or third version of a given format is effectively a new one.[60] Once a format becomes obsolete, it may be difficult or impossible to find and run software that can continue to render it.

Unfortunately, it is not possible to solve this problem simply by saving old rendering software, since the computers on which such old software runs will become obsolete themselves over similarly short timespans. Therefore, the essence of the long-term digital preservation problem is how to maintain the ability to run rendering software for old digital objects.

Emulation versus migration

As noted by Rothenberg (1999) numerous approaches have been suggested to solve this problem, but most of these are (at least in the current state-of-the-art) infeasible and ineffective for the vast majority of digital formats and objects. Of the few that are potentially feasible, only one (emulation) attempts to preserve digital objects in their original form – and to preserve executable objects in executable form. All other approaches translate each original digital object into successive new formats, sometimes initially and, in most cases, repeatedly over time. On the one hand, one advantage of such repeated translation is that future versions of the object are more likely to be in 'vernacular' forms, which are familiar to future users and can be utilised with minimal effort. On the other hand, the disadvantages of repeated translation include the high cost of performing such translations repeatedly over time on every object to be preserved, the likelihood of cumulative corruption via successive translations, and the loss of the authentic original version of each object. Furthermore, the cumulative corruption inherent in successive translation (i.e. migration) undermines the quality of the future vernacular versions of an object, making this supposed advantage of dubious value. Finally, a frequently unrecognised problem with migration approaches is that new formats do not always subsume the functional capabilities of earlier ones, making migration impossible in some cases; this is especially true when paradigm shifts occur, since these often make new formats incompatible with older ones. IBM's UVC-based data preservation approach, which has been explored by KB, avoids many of the problems of migration, although it is still not capable of preserving executable programs or many other inherently digital objects.

Preserving authenticity

In the digital domain, the concept of an 'original' is harder to define than it is for traditional documents. Any copy of a digital artefact that is an exact bitwise duplicate of the original bitstream of that artefact retains all the possible relevant digital attributes of the original. Generally, the medium on which the artefact's bitstream was originally stored is not considered a relevant attribute, since a given bitstream may be stored on any number of different media. Only the bitstream itself is relevant for most purposes. Any preservation technique (e.g. migration) that changes the original bitstream of an artefact in any way poses a threat to its integrity, especially if the original bitstream itself is discarded in the

[60] This is evidenced by the industry-wide policy of not requiring a given vendor to support its own previous program versions prior to the last one or two.

process. However, maintaining the original bitstream of an artefact is not enough to ensure that its integrity is preserved: the correct interpretation of that bitstream is equally crucial. If the original bitstream is not interpreted and rendered in the intended manner, the integrity of the original will be violated.

In some cases – notably those that are the most analogous to traditional page-image printing – the interpretation and rendering of an artefact's bitstream is relatively straightforward. Any reinterpretation or copy (i.e. migration) of such an artefact's original bitstream that produces a page-image that is visually equivalent to the original will retain most of its integrity (up to the limits of the copying process's visual accuracy). However, digital artefacts that rely on the unique capabilities of the digital medium cannot be copied as simple page-images. Such artefacts include any that utilise dynamic, multimedia or interactive content, all of which require that computer programs be executed in order for them to behave correctly. Such inherently digital artefacts cannot be represented as page-images, but must be actively rendered by computer programs executing on computers. This active rendering underlies the ability of such artefacts to exhibit their original dynamic – and potentially interactive – behaviour. Such behaviour may be crucial for enabling future scholars to utilise such artefacts in performing their research. We refer to this as 'behaviour preservation'. Any preservation process which prevents the correct executable rendering of such artefacts (i.e. prevents the preservation of their behaviour) will necessarily sacrifice their integrity, and thereby their authenticity. In many cases, the content itself of a digital artefact does not exist prior to its being rendered: in such cases, it is not just a matter of preserving the behaviour of an artefact but of preserving the artefact at all, in any meaningful sense.

The costs of digital preservation

The paper by Oltmans and Kol (2005) provides a good comparison of the costs of migration and emulation. The Dutch National Archives paper 'Costs of Digital Preservation' (2005) provides independent verification of this same comparison, while other papers from the Dutch National Archives Testbed project provided extensive discussion and analysis of several preservation techniques.[61] Finally, the Life-cycle Information for e-Literature study (LIFE; McLeod *et al* 2006) funded by JISC, has conducted a series of case studies to assess the life-cycle costs of digital preservation.

There are many potential factors involved in the cost of a given technical preservation approach, and different approaches involve different factors. For example, emulation involves the creation of emulators, whereas migration does not; on the other hand, migration involves the repeated translation of individual digital objects into new formats over time, whereas emulation does not. The data preservation approach, based on IBM's UVC, falls somewhere between these two cases.

It is likely that one-time costs, even if they are large, will be swamped by repeated costs. For example, any cost that is incurred for each distinct digital format, which must be understood and reverse-engineered, will quickly overshadow the one-time cost of creating an emulator for a given computing platform. Any per-format cost of this kind, however

[61] Available at: http://www.digitaleduurzaamheid.nl/index.

minimal, will be multiplied many times over the indefinite lifetimes of any digital objects that utilise a given format, as each format is translated into an endless succession of new formats, each of which must be understood and reverse-engineered in turn. Furthermore, this cost will be multiplied by the dozens or hundreds of distinct formats that may be in common use at any given time, Similarly, any per-item cost that is incurred by each individual digital object that must be processed will be multiplied many times over the lifetime of each such object; because the number of objects is typically millions of times greater than the number of formats, this cost is multiplied by a much larger number. In short: for large corpora, any per-format or per-item processing cost is likely to be prohibitive. This is all the more problematic because most objects in any corpus are rarely (if ever) accessed, so any cost that is incurred by preserving them individually will be wasted.

These arguments – along with the two cost studies cited above – suggest that contrary to widespread (yet uninformed) opinion, emulation is likely to be far less costly in the long run than migration, since the latter has significant per-format and per-item costs, whereas these costs are virtually zero for emulation.

For a large digital corpus (which can easily grow to hundreds of millions of documents and beyond), a preservation approach should have essentially a zero per-item processing cost – aside from the inevitable cost of copying the bitstreams of saved objects to new storage media as needed over time. Similarly, per-format costs should be near zero if at all possible. This suggests an approach that treats all formats identically and does no processing at all of individual objects over time.

Moreover, one aspect of cost that is rarely discussed is the cost of losing or corrupting digital objects over time. Any preservation approach has a finite probability of making some objects in its care inaccessible or unusable over time, or of corrupting their content, form, behaviour or appearance so as to make them meaningless or destroy their authenticity. In particular, any preservation approach that discards the original of each digital object in favour of some translated version of that original introduces an inevitable degree of loss, whose cost should be accounted for when that approach is evaluated. Similarly, any approach that successively converts or translates each digital object over time will incur an unavoidable likelihood of corruption, as the cumulative effects of multiple imperfect conversions accrete over time. Finally, any approach that is incapable of preserving executable or other inherently digital objects in executable form will lose the essence of those objects, preserving only the dead husk of a once-living artefact.

Technological approaches to preservation

Despite the community's relative indifference to technological approaches to preservation, these remain the foundation of any long-term digital preservation effort. Therefore, it is useful to discuss some of the relevant aspects of several recent projects which have explored the technical aspects of preservation in depth. However, before doing so, we digress for a brief discussion of the issue of preserving original documents versus vernacular renditions of those original documents.

Originals vs vernacular renditions

For some purposes, originals may be far less convenient and useful than converted, future vernacular renditions of digital artefacts. Such vernacular renditions convert the original into a form that is more familiar to future users. Typically, casual researchers and even scholars use such surrogates (such as modern translations of original texts or photographs of artworks) for much of their research, resorting to originals mainly to validate the surrogates or tease out new aspects of the original that are not represented in the vernacular renditions.

Many proponents of migration argue that it produces the utility of vernacular renditions, since each migrated version of an original is such a rendition. However, migration preserves *only* such renditions, not the originals from which they were derived. This makes it impossible to verify the accuracy of a given rendition. In addition, because the original is discarded, each successive vernacular rendition must be generated from the last one, resulting in cumulative corruption that cannot even be measured, since the original is unavailable for comparison.

In order to produce vernacular renditions, it has been suggested that emulation be designed to facilitate 'vernacular extraction', to help future scholars extract the content of an emulated digital original into a future computing environment in order to produce a vernacular rendition of the original (Rothenberg 2000a; Rothenberg and Slats 2003). Since the original would not be discarded in this case, it would continue to be accessible (under emulation) for the validation of each vernacular rendition, as well as for direct scholarly reference.

UVC data preservation

Originally, IBM's UVC was proposed as a platform for running emulators of obsolete computers. To date, however, the UVC has been applied only to a far less ambitious task, i.e. the future interpretation of encoded versions of saved digital documents. This has been demonstrated for Joint Photographic Experts Group (JPEG) images and spreadsheets, as described in several reports by the Dutch National Archives Digital Preservation Testbed (ICTU 2003). This approach generates a computer-readable logical data description of an original digital document, saving this description with the document itself. In the future, the UVC is implemented on a future computer and the original saved document, along with its logical data description, are read by an interpretation program running on the UVC, which uses the description to decode the original document. This process produces output specifications that are fed into a viewer program (which must be written to run directly on the future computer), which turns these output specifications into actual output on the future computer. Assuming that the format of the original document is understood correctly, that the description is generated from the original document correctly, and that the UVC and viewer programs are implemented correctly on the future computer, this should render the original document correctly on the future computer.

Although this process does require initial processing of each document to produce its logical data description (as well as significant work to understand each format that is to be preserved), it avoids the successive translation of migration and renders the original digital document directly on a future computer.

Furthermore, vernacular extraction using this technique should be easier than using emulation, since the logical data description has already extracted the meaningful content and form of each document, which can serve as the basis for generating future vernacular renditions. The main shortcoming of the UVC is that it cannot preserve executable or interactive behaviour. That is, since it does not serve (at least not yet) as an actual emulation platform, it cannot preserve executable objects in executable form or many other inherently digital objects

Dutch National Archives Testbed (Tessella)

The Dutch National Archives Digital Preservation Testbed project experimented with a range of preservation approaches, including the uses of the UVC and migration. Its detailed reports on each approach continue to be among the most in-depth discussions and comparisons of these techniques that have yet been produced. Unfortunately, although this project initially intended to experiment with hardware emulation as well, that effort was abandoned due to lack of resources.

The overall experience of the Testbed project was that the work required to understand and reverse-engineer each data format was considerable, and that even with concerted effort, some details of each format remain elusive. Therefore, migration and UVC data preservation require much more work than may be warranted for any but the most popular formats. Digital objects that are represented in other, less widespread formats are likely to be orphaned over time, due to lack of resources.

The Erl King

In 2003–2004, the Archiving the Avant Garde project[62] conducted a preservation experiment which involved renewing a 1980s interactive video work, *The Erl King*. The project initially decided to emulate the hardware of the original computer that hosted *The Erl King*, but ultimately this approach was replaced by the creation and use of an interpreter for the source code of *The Erl King* (which was still available). This approach was functionally similar to emulation, in that the original program was run on a modern computer, thereby recreating the exact behaviour of the original work. The result was exhibited in the Guggenheim Museum's *Seeing Double* show and was quite successful (Rothenberg 2006).

KB emulation project

KB's recent emulation project has produced an initial version of a modular emulator (called Dioscuri) for a modern Intel x86 architecture computer. The ultimate goal of this effort is to emulate the full Pentium functionality that is found in KB's Reference Workstation, which is the platform used within the library to provide access to e-journals and other digital holdings (such as CD-ROM publications). The initial version of Dioscuri emulates the equivalent of an 8086 processor, which already is capable of running the MS-DOS operating system and a number of application programs. It is expected that soon this emulator will be able to run the Windows95 operating system along with some preserved applications for that operating system from KB's collection. Eventually, it is hoped that the

[62] See: http://bampfa.berkeley.edu/about_bampfa/avantgarde.html.

full Pentium emulator will be able to run most or all of the code that runs on the Reference Workstation, thereby enabling it to render documents in the e-Depot on machines other than the Reference Workstation.

In order to provide longevity, Dioscuri has been written in Java, which allows it to run on the Java Virtual Machine, which in turn can run on nearly any computer. This would allow the emulator to be used on any future computer that implements the Java Virtual Machine. The project has identified numerous issues already that are relevant to the long-term preservation of digital objects via emulation. Although the work involved in writing an emulator suitable for preservation purposes is significant, it is not large as software development efforts go. No insurmountable obstacles have been uncovered, and the outcome of the project so far suggests that emulation can be a feasible and effective means of preserving digital objects of any kind – including executables and other inherently digital objects.

Migration to a 'future-proof' format

The Extensible Mark-up Language (XML) is frequently proposed as an approach to preserving digital objects, but by itself XML provides only a limited and short-term solution to a small subset of the digital preservation problem. The idea is to use XML to encode every digital object that is to be preserved, using a suitable tagging scheme or Document Type Definition (DTD). For simple page-image objects, such as static text documents or webpages, XML does offer some preservation potential, since it enables such documents to be represented in a form that can be rendered by any web browser or other viewer which can parse the XML structure itself and correctly interpret the tags used in each encoded document. For any document whose original format is not already in XML, this requires pre-emptive migration, in which the original is translated initially into another form (in this case, XML) after which, typically, the original is discarded. Any document whose format is incompatible with the format capabilities of XML and any of its attendant tools (such as Cascading Style Sheets; CSS) must be shoehorned into XML during this initial migration; but in many such cases, the available format options should be sufficient to recreate the form of the original while retaining its content.

However, XML itself does not fully specify the behaviour of any of the tags that it uses to describe format: these tags are defined by communities of interest, and their behaviour depends on the rendering software that interprets them. This is both the strength and weakness of XML: it neither constrains the set of tags to be used in any given XML document, nor specifies the meaning or behaviour of any of those tags – except the ones that define XML's own structure. XML is open-ended, which makes it very flexible; that is, it can be thought of (at least in the preservation context) as a meta-format for describing other formats. But because of this flexibility, it is almost entirely free of semantics. Therefore, the rendering of even the simplest page-image document represented in XML remains dependent on the behaviour of whatever program is used to render it; XML does little to ensure that such documents will be preserved correctly, beyond simplifying the problem by forcing all formats into the single XML meta-format. Furthermore, as with any migration scheme, XML will require additional future migration into some new format when it becomes obsolete – as it inevitably will.

Furthermore, for anything other than simple page-image documents, XML does even less. Like HTML, it provides a framework for representing compound objects, each of whose components may require a different piece of rendering software. For example, references (such as URLs) in an XML document can denote components such as JPEG, TIFF or Portable Network Graphics (PNG) images, spreadsheets, wordprocessing documents, databases or database queries, simulations, animations or the results of executable scripts, server-side Common Gateway Interface (CGI) code, or other interactive programs whose results are intended to be generated and rendered as part of the encompassing digital object. The XML document can point to such external components, but they are not typically embedded within the XML document itself (with rare exceptions), since doing so would make that document huge. Furthermore, many of these components will be in so-called 'binary' form, i.e. unconstrained binary that need not correspond to American Standard Code for Information Interchange (ASCII) or Unicode encodings of text. In any case, each such component must be rendered (and possibly even generated) by a computer program that understands its format – and none of these formats are themselves XML. The encompassing XML document may provide metadata describing such components or specifying what rendering software they require, but aside from this, XML does nothing to solve the problem of preserving and rendering such non-page-image objects in their original form.

Appendix D: 12 e-journal archives compared

This appendix includes some selected tables comparing the 12 e-journal archives included in the survey conducted by Kenney *et al* (2006).

Table 7. Archiving strategy: what type of archiving strategy do the e-journal archives use?

	CSI	ECO	EJC	KB	KOP	LA	LANL	NLA	OSP	PMC	PORT
Migration	•	•	•	•	•	•	•	•		•	•
Emulation				•	•			•			
Normalisation	•						•	•	•	•	•
Reliance on standards	•			•	•		•	•		•	•
Refreshing	•		•	•	•	•[63]		•		•	•
Use of durable media				•	•				•	•	•

CSI = CISTI Csi, ECO = OCLC ECO, EJC = OhioLINK EJC, KOP = Kopal/DDB, LA = LOCKSS, LANL = LANL-RL, NLA = NLA PANDORA, OSP = Ontario Scholars Portal, PMC = PubMed Central, PORT = Portico

Source: Kenney *et al* (2006)

[63]This entry has been adjusted for reasons of accuracy.

Table 8. Documentation type: what kind of written documentation do e-journal archives (plan to) have that explicitly refers to e-journal archiving?

	CSI	ECO	EJC	KB	KOP	LA	LANL	NLA	OSP	PMC	PORT
Mission statement			●	●	●					●	●
Publishers' agreements	●		●	●	●	●		●		●	●
Membership agreements							●				●
Selection or acquisition policies	●		●	P	●	●	●	●	●	●	●
Transfer requirements and deposit guidelines	●		●	●	●					●	
Ingest			●	●	●		●			●	●
Archival storage							●			●	●
Quality control							●			●	●
Auditing							●				●
Data management	●						●			●	●
Disaster planning or recovery	●		●							●	P
Preservation planning					P						P
Metadata				●			●			●	●
Access and use policies							●	●			●
Financial reports											●
Annual reports				●							

P = planned within the next six months. Source: Kenney *et al* (2006)

Table 9. Trigger event: trigger events that spark changes in access for the authorised community

	CSI	ECO	EJC	KB	KOP	LA	LANL	NLA	OSP	PMC	PORT
Publisher ceases operation				●	●	●					●
Publisher no longer offers back issues				●	●	●					●
Copyright expires					●	●		●			
Journal ceases publication				●		●					●
Catastrophic failure				●		●	●				●
Temporary failure				●		●					
Other								●			●

Source: Kenney *et al* (2006)

Table 10. Archiving activity: do the e-journal archives have any relationships with other archiving organisations involving the following activities?

	CSI	ECO	EJC	KB	KOP	LA	LANL	NLA	OSP	PMC	PORT
Exchange ideas and strategies	P		●	●	●	●	●	●		●	●
Share planning documents				●	●	●	●	●			●
Share software			●	●	●	●	●	●		●	●
Coordinate content selection						●		●			
Reciprocal archiving, off-site storage or mirroring	●			P		●				●	P
Secondary archiving responsibility						●					
Shared facilities or resources	●					●		●			●
Other	●										

P = planned within the next six months. Source: Kenney *et al* (2006)

Appendix E: Stakeholders' views and positions on scholarly communication and publishing

In order to understand the issues and developments in scholarly dissemination and publishing, it is useful to understand the different interests, needs and objectives of all the stakeholders involved in these areas. In this Appendix, we discuss the views and positions of a selection of the most important stakeholders. Where relevant, the key findings have been re-iterated in Chapter 3 of this report.

Publishers

As detailed in Section 3.1, STM publishing is a large market with a limited number of large players and a large number of small players. Traditionally, scientific publishers undertake a series of functions that have emerged for efficiency reasons: peer review, copy editing and typography; database preparation; production and distribution; and archiving (Boyce, 1996). A large professionalised service provider can execute these tasks much more efficiently than authors.

There are three main types of STM journal publisher: commercial publishers constitute the largest players in the field (Elsevier, Springer, Blackwell, Wiley, Taylor & Francis, etc.), learned societies and university presses, e.g. Cambridge University Press, Oxford University Press).

Traditional publishing

The traditional business model of publishing is based on license fees for title subscriptions, which changed into package licenses after the Big Deal. The library community often views publishers with scepticism. It seems unfair to them that universities have to 'buy back' their research results. They also argue that the market for STM publishing is, in many ways, an imperfect one. Competition between journals is limited, as no two journals are exact substitutes; they each have a monopoly in a certain niche market and research libraries have weak negotiating power when faced with increased subscription fees. In contrast, publishers argue that publishing has inherent costs and that they add considerable value to the research process by guaranteeing the quality of journal articles (through editorial boards which oversee the academic quality of journals) and distributing them in the most efficient way possible (interview with Paul Ayris 2007; see also Box 2, clause 8 of the 'Brussels Declaration', as shown below). Consequently, publishers rely on the continuing need of university libraries and professional associations for high-quality scholarly papers, whose quality is guaranteed mainly by professionalised, external peer-

review systems. Thus, the business model is built on research institutions' ability to fund and willingness to pay for maintaining library collections.

The sustainability of the traditional model is a subject of intense debate. Some, predominantly traditional publishers, argue that the journal article as a definitive object of research dissemination will continue to dominate scholarly communication. This model will continue to work as long as the peer-reviewed published journal article is the accepted, quality-assured object of academic dissemination.

> *'The reason why the traditional channels are so robust is because one of the parties involved – the (commercial) publisher – is entirely focused on managing the process. The publisher has no stake in the content and no interest other than to produce a journal that has the best possible reputation with its designated readership.'*
>
> Herman Spruijt, International Publishers Association

Additionally, the predominance of academic journals is enforced by research funders, which mostly base their mechanisms for research evaluation on bibliometric analysis of publication and citation volumes. Consequently, high-impact factor journals[64] are considered to provide the highest quality of articles in their specific research field. As long as these mutual interests continue – i.e. as long as authors have an interest in publishing in these journals as a means of developing their reputation and professional careers – the current model will prevail. Publishers are likely to attempt to increase the share of ISI-rated journals in their portfolios, while subscription cancellations are likely to focus on peripheral journals with small audiences. JPMorgan believes that companies such as Elsevier, with strong catalogues of ISI-indexed journals, remain well positioned to outperform the market (as cited by Mark Ware Consulting, 2006). According to this view, publishers will continue to be the main supplier to digital archives, such as KB's e-Depot.

Others argue that these mutual interests may decouple in the coming years with the advent of more informal methods of research dissemination; these views typically stem from the library community and research funders. Potentially, a much more fragmented and diverse supply of scientific outputs may evolve. The development and acceptance of alternative quality assurance mechanisms will be a crucial factor in determining whether or not this occurs.

There are also diverging opinions on the urgency of assuring perpetual access. Generally, publishers claim that there is no access problem: access to scientific information has never been better. Nonetheless, libraries increasingly demand guarantees for perpetual access that go beyond a promise in the license agreement. Additionally, learned societies are beginning to use archiving arrangements with an independent third party as a criterion for selecting a publisher to produce their journals, as publishers begin entering into agreements with long-term archivers. In fact, most of IASTMP's members have signed the so-called Brussels declaration (STM 2007), which includes statements on perpetual access and access to research data (see Box 2, clauses 6 and 7).

---

[64] Journals with articles that are cited relatively often by other papers. The largest database of journal publications and their citations is owned by Thomson Scientific (formerly Thomson ISI), which calculates the journal impact factors. The journals listed in this database are referred to as 'ISI-indexed'.

1. Publishers support the creation of rights-protected archives that preserve scholarship in perpetuity.
2. Raw research data should be made freely available to all researchers. Publishers encourage the public posting of the raw data outputs of research. Sets or sub-sets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars.
3. Publishing in all media has associated costs. Electronic publishing has costs not found in print publishing. The costs to deliver both are higher than print or electronic only. Publishing costs are the same whether funded by supply-side or demand-side models. If readers or their agents (libraries) don't fund publishing, then someone else (e.g. funding bodies, government) must.

Box 2. Clauses 6-8 of the 'Brussels Declaration'

Source: STM (2007)

Open access and self-publishing

Open-access publishing offers an alternative to the traditional publishing model. In the 'author pays' system, subscriptions are free, but the author is charged for copy-editing, peer review and distribution. This charge is usually paid by a research grant or out of institutional funds.

Examples of open-access publishers include BioMed Central (with more than 100 journals in its portfolio) and the journals from the Public Library of Science (PLoS). BioMed Central charges €1,000 per article for most of its journals and PLoS charges €1,200 (European Commission 2007). It is estimated that only 2.6 percent of the core ISI set or approximately 1 percent of the set of journals included in ISI's Web of Knowledge are open-access journals (McVeigh 2004). Similarly, other publishers are beginning to offer open-access options for traditional journals. However, thus far, the level of uptake of open access in hybrid (journals containing 'traditional' as well as 'author pays' articles) journals is relatively low. Typically, the contribution to the total number of papers published is in the range of 2–5 percent (Mark Ware Consulting 2006); in the total output of Elsevier's portfolio, open-access papers constitute significantly less than 1 percent of the total number of papers published (interview with Nick Fowler 2007).

Although open-access publishers generally have to supplement their funding with other sources of income, such as advertising and donations, BioMed and PLoS believe that they will soon break even. However, Butler (2006) reports that PLoS lost US$ 1 million in 2005 and its author fees and advertising revenues covered only 35 percent of total costs. Nonetheless, supporters of open access believe that within a reasonable time, all pure research papers will be open access.

Self-archiving is another element of open access, enabling authors to disseminate their research articles for free over the internet and helping to ensure the preservation of those articles in a rapidly evolving electronic environment (JISC, 2005; House of Commons, 2004). Initially there may be a six-month embargo from when articles first appear in a subscription journal, but within a few years these papers will be open- access available. The

UK research councils, The Wellcome Trust and government-funded research councils have set open-access requirements for papers arising from research that they have funded (The Wellcome Trust 2006). Similar shifts are evident in Canada and France. NIH funding of research generates 60,000 articles a year (NIH 2003) and NIH is expected to shift to an open-access policy (interview with Robert Kiley 2007). It will probably require open-access availability of articles with a somewhat longer embargo period.

Often, open access is viewed by research libraries and funders as more sustainable than traditional publishing, since it can be linked more directly to research funding; dissemination can become a distinct component of a research grant. Self-archiving or institutional open-access repositories such as PubMed (NIH), UK PubMed (Wellcome Trust) and university digital repositories could potentially harm publishers' revenues in the longer term, although they have not yet made a big impact, according to traditional publishers.

National libraries

National libraries have a public responsibility to collect published information, preserve it and provide permanent access to it for research, education or any other purpose in society. Libraries often have a long history of archiving and preservation; KB was established more than 200 years ago from the original collection of William of Orange; the US Library of Congress was established by an Act of Congress in 1800; and the Bibliothèque Nationale de France traces its origin to the royal library founded at the Louvre by Charles V in 1368.

Various countries have specified a legal deposit function which obliges publishers to deposit their publications in one or more national repositories. Recently, national libraries are beginning to acknowledge that, due to the increasing proportion of electronic publishing in scholarly communication and in order to continue fulfilling their public responsibility, they should invest in digital archiving services. National libraries rely on their governments' willingness to shoulder part or all of the required investments and costs. Once the investment has been made, these services should be scalable to include a wider range of content and provide value-added services that may generate additional revenues and/or strengthen the national libraries' public good function. For example, traditionally, the British Library has been able to generate substantial revenues through its Document Supply Service, run from Boston Spa.

Other stakeholders' perceptions of national libraries vary considerably. Often, libraries' long history is viewed as an important asset, creating confidence that they will continue to exist in perpetuity. Furthermore, their not-for-profit status and public role avoid suspicion of underlying economic motives. The public interest is best served by libraries to ensure all publications are collected and preserved in perpetuity. There is no apparent commercial or other reason why a for-profit organisation would incur cost to guarantee longevity of data and continued accessibility if there is no financial reward for doing so.

However, there are criticisms of the role of national libraries. For example, sometimes the requirements imposed on publishers to conform to a national library's legal deposit of publications are met with frustration (interview with Andy Williams 2007). This is particularly true for publishers faced with a legal obligation to deposit their content with a national library, such as in the UK. Additionally, the British Library has associated itself

with the open-access movement through its active involvement with UK PubMedCentral, thus giving the appearance of favouring a particular business model. This has lead to irritation among some publishers (interview with Pieter Bolman 2007). In The Netherlands, the distinction between preservation and access is not always clear. KB is facilitating on-site access to its e-Depot content for its 'walk-in' users and does not require a separate license agreement for such access. Herewith, KB has come to shift towards facilitating research in other areas than its historic remit. Whereas traditionally, KB focused on the humanities and social sciences, it has moved slowly into the territory of traditional science and technology libraries (e.g. the Delft University of Technology Library). This is perceived by some libraries as an ambiguous situation (interview with Maria Heijne 2007).

Furthermore, despite KB's solid reputation as a front runner in R&D, the role and remit of KB is largely unknown outside The Netherlands. Being funded by the country's taxpayers and having a 'national library' stamp also may create the perception that KB will act only in the national interest; this may work against the library's aspirations in the global arena, where national boundaries are becoming less relevant.

Research libraries

In the broadest sense, academic libraries provide information services to their academic researchers (sometimes referred to as 'clients'). In order to be able to 'stand on the shoulders of giants',[65] researchers need to have access to the work of these giants. Traditionally, formal, peer-reviewed publication has served to support the tenure and promotion process, as well as providing the formal record of progress in science, but some observers see both of these roles as being usurped to some extent by informal publication.

Often, umbrella organisations for consortia of universities negotiate collective license agreements with publishers; for example: SURF (The Netherlands), JISC (UK), Deutsche Forschungssgemeinschaft (Germany) and the Danish Library Agency (Denmark).

As reported by the House of Commons Science and Technology Committee (2004), libraries are atypical consumers. Rather than purchasing more goods until the benefit they receive is balanced by the cost, they spend up to the limit of their budgets. If prices rise, libraries will purchase fewer journals; if prices fall they will purchase more. Similarly, if research output rises but library budgets remain unchanged, libraries will cancel the least popular journals. The ceiling on budgets means that publishers of 'must-have' journals have a monopoly position and can increase their revenues when they raise prices, as lesser journals are discarded by libraries. This is why observers from university libraries expressed little confidence in the motives of large publishers.

Researchers and authors

Researchers and authors constitute the main demand for journals, via their libraries. However, they are becoming less willing to visit a physical library. Across academic fields,

[65] Isaac Newton's modest reflection on his academic achievements, when famously writing in a letter to Robert Hooke in February 1676: "If I have seen further it is by standing on the shoulders of giants".

scholars have come to value electronic access to published research output (Kenney *et al* 2006) and the increasing use of electronic scholarly literature correlates with its online availability (Guthrie and Schonfeld 2004; Tenopir 2003). Increasingly, researchers want the information delivered to their desktop computers, in order to be able to use it in the lab, at conferences, in a hospital, etc. Researchers and authors, especially those in academic posts, have relied on peer-reviewed journals from traditional publishers to provide opportunities for publication and dissemination of quality research findings, on which their professional reputations and careers depend. Open-access publishing and new forms of quality assurance are emerging as alternatives to traditional publishing. There have been numerous studies of author behaviour, perception and attitudes. The results of two recent large-scale surveys are summarised in a recent White Paper on scientific publishing (Mark Ware Consulting 2006). The combined conclusions of these surveys conducted by the Centre for Information Behaviour and the Evaluation of Research (CIBER) (Mabe 2006; Rowlands *et al* 2004, 2005) can be summarised as follows.

- Although the most important reason given for publishing was to disseminate the results, the underlying drivers were funding and furthering the author's career.
- In choosing where to publish, being able to retain copyright or being able to place a copy of the pre- or post-print material on the web or in a repository, were not of importance to most authors.
- The importance of peer review is underlined. There was near-universal belief that refereed journals were required.
- Reading patterns are slowly changing: a significant minority (22 percent) of respondents preferred to conduct their e-browsing from the comfort of their home.
- There was high demand for articles published more than 10 years ago.
- The awareness of open-access publishing is increasing and more authors claim to have published in an open-access journal (this grew from 11 percent in 2004 to 29 percent in 2005).
- Authors had very little knowledge of institutional repositories.

Other stakeholders

Academic or learned societies

The mission of academic or learned societies is to advance knowledge in a particular field through improving research networks and dissemination of research. Many leading academic journals are edited and published by learned societies.[66] Societies generally tender their journals to publishers every few years. This gives them potentially significant influence over publishers' future commitments to archiving. Journal migration from publisher to publisher causes considerable irritation in the library community.

[66] See the Association of Professional and Learned Society Publishers website: http://www.aplsp.org.

Research funders

Research funders have an important role in scholarly dissemination. After all, it is they who wield the financial resources needed to conduct the research process, and publishing results is but one aspect of this process. Observers have reported that an increasing number of research funders advocate open-access publishing and other types of information sharing, thereby improving access to research information. Funders will continue to demand that research outputs be captured, stored and preserved for long-term access. As an example of research funders' increasing influence on how research results are disseminated, it is useful to mention that British research councils have adopted a new policy that requires the researchers they support to publish their findings in peer-reviewed, open-access journals.

Governments and regulators

Governments depend on safe storage, preservation and access to data and see this as a public service that they will invest in to develop the capability. Governments also want to ensure wide access to knowledge, and support effective generation of high-quality academic output in order to realise the economic benefits from innovation and commercialisation of these outputs. Intellectual property rights are intended to create an incentive for the generation of scientific knowledge.

The current VAT regime in Europe makes electronic-only delivery of information services (i.e. STM publications) more expensive than paper-based delivery, since often the VAT for paper publications is less than that for digital objects (Dewatripont *et al* 2007). Perversely, this incentivises paper and undermines the switch to e-only publication.

Intermediaries

As reported by Versita (2006), there are also different types of intermediaries, such as gateways, aggregators and content hosts. The gateway model has been adopted by almost all of the major subscription agents in the world in the electronic journal market. The gateway is a large collection of links to publishers' full text content. Gateways do not host or own copyright to the full text, but they accumulate information about the full text. Full-text aggregators create databases of full text articles, defined by subject area and sold as a single product, rather than as individual subscriptions to components of the database (for example, EBSCO Publishing, ProQuest and Ovid). Content hosts are organisations which provide a hosting service for publishers including data conversion, secure online hosting and distribution of material to subscribers and pay-per-view customers. Examples include Ingenta, HighWire Press and MetaPress.

Appendix F: Stakeholders' views and positions on preservation

This appendix provides a brief summary of the different perspectives on preservation through a stakeholder-by-stakeholder discussion. The sections also elaborate on how these stakeholders are viewed by other players in the arena. Where relevant, the key findings have been re-iterated in Chapter 4 of this report.

University libraries

Traditionally, university libraries function as information hubs for their academic communities, delivering not only copies of their existing collections but also articles from a long list of subscriptions. With the emergence of the electronic age, their role has become much more fluid. The important question that will be answered in the coming years is whether they will continue their function in rejuvenated form as information nodes for scholarly work, or whether they will be relegated to the position of safe keeper of historic paper documents, most of which eventually will be remotely accessible by means of digitisation projects (e.g. Google and Microsoft).

At this moment, many university libraries are working hard on developing and/or implementing strategies aimed at achieving sustainable business positions as information nodes in smaller or larger academic communities. These often involve alliances with other university libraries or third parties. Archiving and preservation are regarded often as a cornerstone of these strategies, which means that university libraries will have to decide what their position should be regarding national libraries and other stakeholders.

Perpetual access concerns

As discussed in Section 3.2, research librarians are increasingly concerned with perpetual access. Since a shift to electronic-only material seems inevitable, libraries need to prepare for a future in which they do not have academic journals stored in a print archive. If, for example in a worst case scenario,[67] access to the *British Medical Journal* or *The Lancet* were disrupted, this would strongly affect a university department whose core business is medical research (interview with Paul Ayris 2007). Therefore, when negotiating new license agreements, clauses on arrangements with third-party digital archives become particularly important.

[67] The probability of this scenario is seen as very low.

> *'Our university has an internal policy to only sign licenses with publishers who have some kind of perpetual access arrangement in place.'*
>
> Maria Heijne, Delft University of Technology

One observer commented that if journal license agreements contain explicit phrases on perpetual access – specifying that publishers are responsible for providing access for universities in case of a trigger event – archiving costs should be included already in the license fees (interviews with Matthew Cockerill and Kurt de Belder 2007). Organisations such as JISC are in a good position to insist on these terms.

Some observers from the library community expressed scepticism about preservation initiatives which involved a national library or government body. It was argued that these national library initiatives are heavily dependent on government funding, and that due to their national stance, they may not always act in the interest of the international community.

Authenticity

Librarians seemed concerned with the dangers of intentional human intervention, whether resulting from monetary, intellectual, ideological or political motives. Here again, national libraries seem to be among the most trusted of such institutions, since they are perceived as having the fewest potential conflicts of interest. However, several observers noted that even governments have not always been above attempting to modify the scholarly record; so the best arrangement would be a consortium of national libraries that maintain replications of each other's repositories and conduct periodic cross-audits of each other's holdings ('trust but verify').

Institutional repositories

In parallel to relying on a third-party digital archive as an insurance policy for perpetual access, large universities have begun to host publishers' back files on their own servers, which means creating institutional repositories. Most universities do not have the resources to do this. Therefore, umbrella organisations in Europe (SURF, JISC, etc.) are collaborating to set this up. For example, Dutch universities have created a network of institutional repositories (DAREnet) for publishing their research results; KB provides the long-term digital archiving and preservation for DAREnet through a formal agreement. In a broader context, European universities express the desire to create a similar collaborative infrastructure of institutional repositories. The European Commission would support a European version of the DARE initiative.

Long-term preservation

With a few exceptions, most institutions and scholarly communities appear to be ignoring long-term preservation issues, focusing instead on very short-term solutions which may last no longer than five or at most 10 years (interview with David Prosser 2007). This appears to emanate from a belief that the long-term problems are intractable, so worrying about them at this time will prevent making any progress. However, some argue that if these problems are not faced upfront, any short-term approaches that are adopted are likely to fall far short of future needs.

> *'While technicians are warning of possible nightmare scenarios for technical obsolescence, librarians do not yet have the full awareness of the future problems of digital preservation.'*
>
> A Librarian

Within the scholarly community, there is considerable scepticism about the preservation of scholarly communication. First, several observers commented that in order for any organisation to be a credible candidate for performing preservation, preservation must be part of its core mission. Because traditionally, publishers have not considered preservation to be part of their business models, they are considered by many to be unlikely to do a good job of preserving scholarly material. The general perception is that publishers may be coerced into preserving material, for example, to guarantee future access to e-journals that they publish, but preservation holds little positive value for them – only the avoidance of negative value (i.e. avoiding litigation by subscribers who can no longer access back issues of an e-journal). Even the possibility of new long-tail revenue is not seen as sufficient to make preservation part of the core missions of most publishers.

Yet another aspect of research libraries' confidence in a digital archive is an organisation's credentials as a steward for scholarly communication. This involves the ability to provide long-term assurance that the authenticity of the scholarly record will be maintained. One observer noted that: "In order to stand on the shoulders of giants, one must know whose shoulders they are."

Finally, several observers expressed concern that key preservation functions should not be outsourced. Even if preservation is a core mission for an organisation, it may decide to outsource some aspects of the process. This is seen as a questionable strategy, since it may reduce the organisation's intellectual investment in preservation or reduce its competence in performing the process. For a library, outsourcing preservation is seen as tantamount to outsourcing the library's collection, which is seen as its core asset, and therefore something which should not be entrusted to others.

This scepticism about the motives, credibilities and competencies of various organisations seems to overshadow technological issues for most observers. Organisational and business model concerns seem paramount to many in the community, with technology a distant second. This may be due partly to the fact that few observers are thinking about long-term preservation. Instead, they are focusing on very short timeframes of no longer than about five years, during which time technological issues probably can be ignored, especially with regard to page-image artefacts such as e-journals.

National libraries

In order to fulfil their public responsibility as the national (legal) deposit, national libraries need permanent government funding. According to most observers, it is plausible for each major economy to have a national institution taking care of national research output. This role is regarded as having paid off when publishers were digitising their back files and making them available online: the legal deposit libraries often had to support these initiatives by retrieving missing volumes. Moreover, now that national boundaries are disappearing in scholarly publishing, it may be more cumbersome for national libraries to select publications with national relevance than to simply archive everything.

National libraries may be classified into three categories. A number of libraries, including the British Library, Deutsche Nationalbibliothek, the national Libraries of Australia and New Zealand and KB are taking a proactive role in electronic archiving and preservation. A small number of libraries specifically have included electronic archiving and preservation in their mission and strategy but have not made substantial progress yet in implementing these intentions. In this group we can include the US Library of Congress. The third, large group consists of libraries that have not made plans or taken steps yet towards electronic archiving and preservation.

Despite their clear role in archiving and preserving national scholarly communications and cultural heritage material, the national libraries' remit is not always clear to stakeholders. Research universities in particular question the roles of national libraries.

> *'National libraries' remit is not always clear; is KB a national public library, a research library, or does it facilitate services to research libraries?'*
>
> Kurt de Belder, Leiden University

National libraries appear to be among the most trusted candidates for preservation, although among these, some (including KB) seem to be more respected than others. Furthermore, the national policies that guide any such institution must be scrutinised. In the perception of some observers, there seem to be differences between KB and its Anglo-Saxon counterparts. The role of KB in the upkeep of these international publications is met with some scepticism from Anglo-Saxon organisations, including universities and publishers, which have a tradition of focusing on initiatives from within the Anglo-Saxon world. The national library of a non-Anglo-Saxon country is not immediately an obvious partner for this task. Many observers posed the question of what would happen if the Dutch Government lost its interest in e-Depot, or was forced to cut funding.

> *'The e-Depot is a great gift from the Dutch Government and the Dutch taxpayer to the international academic community. But can we expect the Dutch Government to continue funding such archiving services of research material for the entire world into the indefinite future?'*
>
> Paul Ayris, University College London

Ultimately, a consortium of national libraries, possibly allied with other institutions, seems to have the most credible motives of any organisational entity. Although such consortia seem to be viewed with favour, some observers noted that KB should not consider an organisation such as Portico to be its peer.

Publishers

Most publishers acknowledge that their industry has a collective responsibility for preservation (see Box 2, clause 6 of the 'Brussels Declaration'). Other publishers see preservation as a public responsibility. Publications deliver commercial returns to publishers over a relatively short period, so older material is commercially less interesting to them. However, technical obsolescence may enable them to repackage and re-sell their content.

Nevertheless, if publishers commit to perpetual access for e-publications, logically they cannot ignore preservation and archiving questions. There are coincidental synergies between publishers and those interested in preservation because, for example, a generic, independent format could enable publishers to distribute their content over various channels onto various platforms: print, monitor, personal digital assistant (PDA), etc. However, long-term preservation has never been a priority for publishers, and it is not something that editorial boards have been very concerned about (although this attitude seems to be changing gradually). Licensing agreements use expressions such as 'an effort to preserve'. Some commercial publishers are signing agreements with digital archives such as KB e-Depot, Portico and CLOCKSS to satisfy customers (i.e. research libraries). But libraries do not have sufficient negotiating power yet to enforce preservation agreements by publishers.

The publishers that do have archiving agreements can convince university libraries more easily to switch to electronic-only journals, which is a more profitable business model than print publishing. Also, they can offer this to learned societies which want their journals published. However, publishers are wary of archiving agreements that involve significant compliance costs (e.g. that specify the formats of material for ingestion) or opportunity costs (i.e. loss of sales).

Those publishers least likely to go out of business – the traditionally large ones – are those participating in multiple preservation systems. Small and less well-established publishers (or university units or academic societies that publish one journal) are the most vulnerable group in terms of preservation and perpetual access. They are less likely to participate in any of the preservation initiatives. They would prefer these archives to be 'dark', instead of providing access. It will be difficult to convince all publishers to sign up with more than one repository.

The commercial publishers may want three to five depositories to achieve preservation of their full portfolio internationally: at least one each in Asia, Europe and the USA. Or they may achieve sufficient coverage through a single depository with different local mirrors (copies). Multiple repositories may mean multiplication of costs, especially if preservation is conducted in different ways; however, this may provide added insurance, since alternative preservation methods can back-up each other.

Some publishers have indicated that they are not comfortable with agreements that grant access to the whole world in case of a trigger event. According to one observer representing a publisher: "If the content is opened to the whole world without restriction, there is a real risk of piracy and once the trigger event is over, control of the content will be lost" (interview with Steven Hall 2007). In contrast, one publisher mentioned that granting temporary out-of-license access to licensed content after a trigger event could be a useful marketing tool to advertise journals (interview with Peter Hendriks 2007).

In the current open-access environment, there is not necessarily a clear view on the role of preservation into perpetuity. Some observers argue that open-access proponents seem to expect that anything that is accessible will be preserved automatically. Although there are some exceptions, there is yet no coherent plan for preservation agreements in the open-access environment (interview with Nick Fowler 2007).

Some universities which have institutional repositories acknowledge that they have not thought about preservation very much. Their promises that the content of the institutional archive will be preserved are not based on substantive arrangements for this. When putting in place appropriate arrangements for preservation, open-access repositories do not need complicated mechanisms managing access rights of licensees.

Governments

Governments support libraries, research and teaching institutions to provide broad public and professional access to information, as part of their overall national and international economic, educational and cultural policies. Governments ensure:

- that appropriate national and international standards govern legal deposit;
- that appropriate regulations and policies govern freedom of information and data protection;
- compliance with those regulations and policies by each government's own archives and collections of official records.

Government funding to national libraries is motivated by a desire to support and maintain national culture and heritage. Accordingly, most countries have laws and/or regulations obliging local publishers to share their output with the national library and obliging the national library to collect and archive these publications. To date, in most instances this has resulted in rather limited support for the development and use of new electronic archiving and preservation technologies. The exceptions are:

- KB – where the Dutch Government has decided to support its e-Depot venture directed at all English-written articles in STM;
- the British Library – which has a mission to collect all English-speaking publications; and
- the Library of Congress – which has a national mission directed at what currently is the largest scholarly nation in terms of output volume.

Other examples include the national Libraries of Australia and New Zealand. However, in the future, governments' funding priorities for libraries and archives may change.

Other stakeholders

Research funders

Research funders seek returns (economic benefits, international competitiveness, reputation) for investing in high-quality research. They want published articles as evidence of the outputs of research that they have funded, citations as evidence of the uses and outcomes of that research, and assured access to relevant research and other information at zero or low cost, for their own research, teaching and learning. Assured access means rapid, perpetual access, including active links to related material and, where relevant, the authentic intellectual content of the sources.

Scholars

Scholars have roles as authors as well as users of disseminated results. Nowadays, in their role as authors, scholars have multiple means at their disposal to disseminate their findings.

Apart from the traditional publishing route, they can subscribe to an open-access publishing channel, communicate with their peers in more informal ways (e.g. using blogs and wikis) and circulate their articles in the form of 'grey literature'. Observers point out that generally, researchers are not so concerned with preservation issues. The issues that concern them most are perpetual access and tenure and promotion rather than technical obsolescence, particularly in the humanities.

Learned societies

The mission of learned societies is to advance knowledge in a particular field by improving research networks and disseminating research. Societies generally tender their journals every three years and these journals frequently migrate between publishers. This poses difficulties for the production departments of publishers and irritation for licensees. For societies, it is an asset if publishers have preservation agreements, since it is in a society's interest to have the content of its journals available for future generations. But it is taking time to get learned societies on board with preservation.

Systems developers

Over the past years, system developers gradually have gained an (economic) interest in developing solutions or products (hardware, software, operational procedures) to enable digital archiving. Generally, they partner with a launching customer, which could lead to agreements with promising digital archive schemes in the future. Current systems are custom designs and the marketability of these solutions appears limited, but this may change as new applications are developed and the value of stored data is increased. Owners of repositories may decide to take sole or shared ownership of their systems, or leave this to the system developer, depending on its philosophy of knowledge-sharing and ability to upgrade the system continuously. Ownership of the system by the repository may increase dependency on one technology. The DIAS system (used by KB's e-Depot) uses as many standard components as possible (e.g. DB2, a database management system) to avoid the cost of maintaining a highly-customised system.