RAND EUROPE

This PDF document was made available from www.rand.org as a public service of the RAND Corporation.

Jump down to document ▼

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.

Support RAND

Browse Books & Publications

Make a charitable contribution

For More Information

Visit RAND at www.rand.org

Explore RAND Europe

View document details

Limited Electronic Distribution Rights

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

RAND EUROPE

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

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.
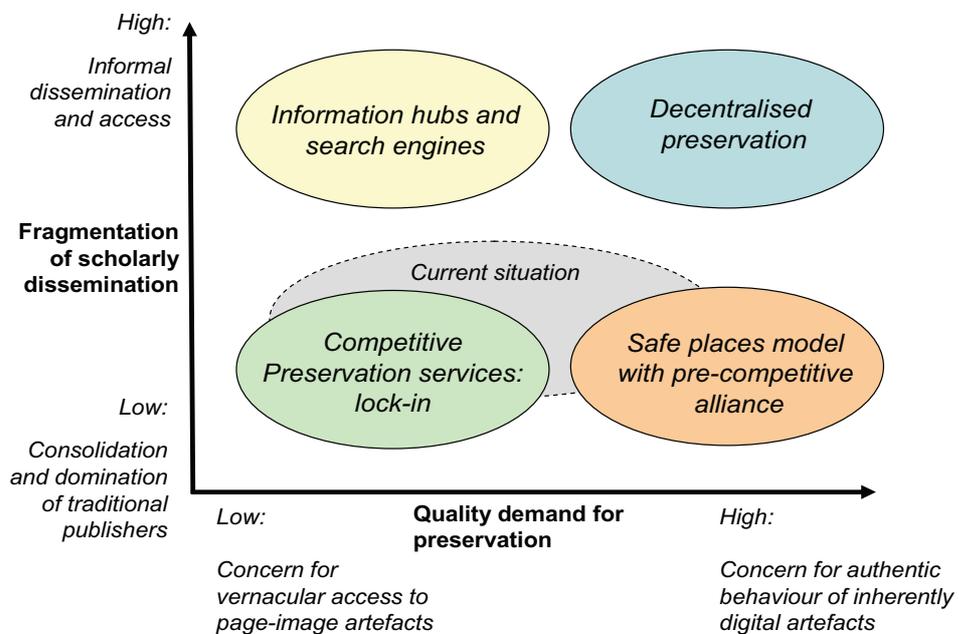

Figure A: Scenario framework for the future of digital archiving and preservation

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.