Preserving the outputs of scholarly communication for the long term: a review of recent developments in digital preservation for electronic journal content

Michael Day[1]

Digital Curation Centre, UKOLN, University of Bath, Bath BA2 7AY, United Kingdom
m.day@ukoln.ac.uk

Introduction

Since their origins in the seventeenth century, scientific journals have become an essential part of the process of science and scholarship. The scientific literature is cumulative, enabling researchers to build upon the work of those that have gone before them through acknowledgment and citation. John Ziman has noted that the citation of references validates many of the claims made in published papers and embeds them in the pre-existing consensus [1]. Until very recently, research and national libraries took most responsibility for the long-term stewardship of this part of the scientific record, working collectively to ensure continued access to the content of printed journals. While this system was not perfect in every single way, its success in preserving content of value was based upon distribution and redundancy. Dale Flecker has pointed out that in the print era, libraries subscribed to and maintained large and highly redundant collections of journal content, also investing in a range of activities intended to maintain usability but which also actively supported their long-term preservation [2]. As elaborated by Sadie Honey, "since multiple libraries subscribe to and process the same journals, there is a high-probability that at least one copy, if not multiple copies, of each issue of those journals will be available for future scholars" [3]. In the digital environment, however, all this has changed.

Research projects in the 1980s first proved that electronic journals were feasible. In the last years of that decade, journal publishers began to experiment with delivering journal content through online networks, starting with initiatives like ADONIS and the American Chemical Society's Chemical Journals Online service. However, it was the emergence of the Internet as a mass medium in the early 1990s that acted as a catalyst for the widespread adoption of electronic publishing methods by journal publishers. Initially, the use of technology was conservative; creating online services that in the majority of cases provided parallel access to journals that were usually also available in printed form [4]. Over time, however, many new features have been added to the electronic versions of journals, meaning that publishers increasingly treat them as the versions of record [5]. In addition, in order to meet user demands and to save costs, many libraries are now beginning to cancel print subscriptions in favour of

[1] Draft version (October 2006) of a chapter prepared for: *Ejournals Management and Access*, ed. Wayne Jones (Binghamton, N.Y.: Haworth Information Press, forthcoming 2007).

licensed access to the online versions. This means that the traditional role of libraries as the custodian of journal content is increasingly uncertain.

At the heart of this problem is the fact that in the digital world, libraries and other institutional subscribers no longer tend to purchase content outright. In the digital era, libraries tend to sign agreements (contracts or licenses) with journal publishers or aggregators that enable authorized users to access digital content hosted elsewhere for a particular period of time. As Ann Okerson noted over ten years ago, the move to licensing models means that subscribing institutions no longer physically own the content that they are paying for, potentially meaning that if, at the end of the licensing period, "they cease paying the lease price, prior investment may become worthless if the information is taken away" [6].

Licenses have two main consequences. The first relates to Okerson's observation that e-journal subscribers have no guarantee that content that has been paid for will continue to be available once the subscription is cancelled. When a print subscription is cancelled, the subscribing organisation does not need to return the back runs of the journal to the publisher. On the other hand, if a license is terminated, continued end-user access to older content can be at the discretion of the publisher. The answer to this 'perpetual access' problem lies in better licensing regimes. As a consequence, many existing e-journal licenses do include provisions for enabling some kind of continued access to content that was previously subscribed to. For example, the current model license developed for the UK higher and further education and research communities - the Model NESLi2 License for Journals - includes an obligation on the participating publisher to provide licensees with perpetual access at no charge to the full-text of purchased journals on termination of the license, either through continued online availability or by the supply of archival copies to the institution or a central facility [7]. Many other national site licensing initiatives, library consortia and individual institutions include similar provisions in their license agreements with publishers [8]. It is clear that enabling perpetual access to content is an important issue that will need further consideration as libraries increasingly drop their print subscriptions in favour of online access to e-journals.

While better licenses can help with solving the problem of perpetual access, the licensing of e-journal content has a second consequence is far more difficult to solve. We have already mentioned that in the print era, the long-term preservation of the scientific record depended upon the distribution and redundancy inherent in the global library system. In the current era of licenses, however, ownership of and responsibility for the preservation of content remains with publishers. While it will not be in the commercial interest of publishers to deliberately destroy content, the fact that it is managed by a single organisation would appear to make it more vulnerable than was the case for printed journals [9]. This deeper problem has been outlined in a statement resulting from a meeting held in New York to discuss the preservation of electronic journal content in September 2005 [10]:

> *Although some - but certainly not all - licenses now recognize that libraries have permanent rights to use electronic journal content, these rights remain largely theoretical. If a publisher fails to maintain its archive, goes out of business or, for other reasons, stops making available the journal on which scholarship in a particular*

> *field depends, there are no practical means in place for libraries to exercise their permanent usage rights and the scholarly records represented by that journal would likely be lost.*

For these reasons, publishers and libraries have begun to seek mutual co-operation on ensuring the long-term preservation of e-journal content. Examples of this are the electronic archiving agreements that the National Library of the Netherlands (Koninklijke Bibliotheek) has signed with Elsevier Science, Springer, and a number of other journal publishers since 2002 [11]. Publisher and library co-operation also underlies the business model that underlies the Portico e-journal archiving service launched in 2005. We will introduce these initiatives in more detail later in this chapter.

The remainder of this chapter will investigate the long-term preservation of e-journal content in more detail. First it will explain why digital materials are difficult to preserver and look at some of the main solutions that have been proposed to date. Secondly it will introduce a range of initiatives specifically related to the preservation of e-journal content, including the different preservation models offered by Portico and LOCKSS (Lots of Copies Keeps Stuff Safe) as well as the e-Depot run by the Koninklijke Bibliotheek (KB) and PubMed Central. Thirdly, it will briefly look at some of the wider problems of preserving scholarly communication in the digital era, focusing on e-print repositories, research data and Internet references.

Defining the digital preservation problem

Digital preservation can be understood as referring to the whole range of activities that are required to ensure that digital objects remain accessible for as long as they are needed. In a much-cited definition, Margaret Hedstrom says that digital preservation involves "the planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable" [12]. Despite the growing ubiquity of digital information, the long-term preservation of information in digital form is far from a simple task. At the heart of the problem is the rapid obsolescence of the various technologies on which digital information depends, as outlined in the highly influential 1996 report of a task force set up by the Commission on Preservation and Access (CPA) and the Research Libraries Group (RLG). The group noted that "rapid changes in the means of recording information, in the formats for storage, and in the technologies for use threaten to render the life of information in the digital age as, to borrow a phrase from Hobbes, "nasty, brutish and short" [13]. In addition, digital information is very easy to manipulate, meaning that it can easily become corrupted, whether deliberately or accidentally [14]. Future users of digital resources need to have confidence that preserved objects are authentic in that they are what they claim to be and that their integrity has not been compromised. While there are technical methods available for dealing with this issue at the bit-level (e.g., using cryptographic techniques), confidence in an object's authenticity will ultimately be based on the level of trust a user has in the organization responsible for preserving it. Another set of challenges relate to the legal contexts of digital preservation. So, for example, intellectual property rights (IPR)

legislation or overly restrictive licensing regimes can sometimes restrict the collecting and preservation activities of research libraries and other cultural heritage organizations. Indeed, Alexandre López and Charles Oppenheim have noted that recent changes in IPR law have tilted the balance of rights away from users in favour of content owners [15]. While for some national libraries, carefully constructed legal deposit legislation can help to solve some of these challenges; many of the technical strategies proposed for solving digital preservation problems depend on the adaptation (or re-engineering) of application programs in ways that would not be permitted by typical software (or content) licenses.

As Hedstrom's initial definition suggests, the challenges of digital preservation are multifaceted, involving a mixture of technical and organizational issues. Successful solutions will depend upon what Abby Smith describes as the "series of actions that individuals and institutions take to ensure that a given resource will be accessible for use at some unknown time" [16]. The following section will introduce the most important of these.

Solving the digital preservation problem

Over the past decade there has been steady progress in development of responses to the digital preservation problem, not least in the advocacy of a number of different technical approaches to preservation and a growing recognition of the importance of metadata. This section will outline some of these developments in more detail, focusing on four main topics: the significant properties of objects, the development of repository models and preservation strategies, and emerging standards for preservation metadata and content packaging.

Determining the significant properties of objects

Most digital objects are inherently complex. For example, Kenneth Thibodeau suggests that digital objects inherit properties from three different object classes [17]:

> *Every digital object is a physical object, a logical object, and a conceptual object, and its properties at each of those levels can be significantly different. A* physical *object is simply an inscription of signs on some physical medium. A* logical *object is an object that is recognized and processed by software. The* conceptual *object is the object as it is recognized and understood by a person, or in some cases recognized and processed by a computer application capable of executing business transactions.*

The complexity of the relationships between these object classes means that those responsible for preservation need to make important decisions about which particular properties (or characteristics) need to be maintained over time. In the digital preservation literature, these are often referred to as significant properties. To simplify somewhat, those preserving text objects might need to consider the relative importance of preserving features like layout, fonts, spacing, pagination or colour. Those preserving

images will need to evaluate the importance of features like image resolution or colour. Understanding the significant properties of objects is extremely important in the digital environment because many preservation strategies depend on the periodic transformation (or normalization) of objects or on the development of tools that emulate the behaviour of obsolete hardware and software. It can also be extremely difficult, in part because those responsible for preservation need to have a detailed understanding of what future users might need [18]. It can also be very difficult to be completely objective about significant properties. Hedstrom and Christopher Lee have noted that definitions "of significant properties that affect the aesthetics, implied meaning, and affordances of digital objects tend to be ... subjective and tied to the context of creation and use" [19]. Despite this, determining the significant properties of objects will be a vitally important part of any response to the digital preservation problem.

The relatively limited number of delivery formats used by e-journal publishers at the present time simplifies to some extent the determination of significant properties in the e-journal context. After initial experimentation with simple formats like plain text and bit-mapped images, e-journal publishers have for the most part settled on delivering journal content in two main ways - often in parallel [20]. The first of these is Adobe's Portable Document Format (PDF), which retains many of the features of the traditional printed product and is widely used where an electronic version of the journal is made available in parallel with a printed version. The second main way of delivering e-journal content is through structured formats like the HyperText Markup Language (HTML) and the Extensible Markup Language (XML). HTML is popular as a delivery format - at least for abstracts and reference lists - because journals can take advantage of the hypertext features available in Web browsers. Many of the bigger publishers now store most of their e-journal content in an internal format based on XML or SGML (Standard Generalised Markup Language) and convert this into PDF and HTML for delivery to end-users [21]. Those with responsibility for preserving e-journal content will need to determine which format should be the main foci of preservation, and at least whether it should be the 'added-value' internal source files held by the publisher, or the derivative versions delivered to end-users through publisher or aggregator portals. Focusing on the former is likely to require additional negotiation with publishers or other content owners. A number of e-journal preservation initiatives have decided to focus on publishers' source files, transforming these into a standardised XML-based format - most often the National Library of Medicine (NLM) Archiving and Interchange DTD.

Other types of e-journal content may be more difficult to deal with. Flecker mentions the types of 'supplementary materials' that increasingly accompany journal papers [22]:

> *[These include] files containing detailed research data, further explication of the article information, or demonstrations of points made in the article. These files contain many types of information (statistical data, instrumentation data, computer models, visualizations, spreadsheets, digital images, sound, or video) and come in a wide range of formats, usually dependent on whatever technical tools the author is using at a given moment. Journal editors and*

publishers frequently exercise no control over these formats, accepting whatever the author chooses to deposit.

To complicate matters further, there is the secondary question about what should happen to publishers' delivery services like SpringerLink or Elsevier's ScienceDirect. While these are not themselves part of the scientific record, there may be some perceived value in preserving at least some aspects of their functionality or look-and-feel. Considering this matter seriously takes us into the realm of Web archiving initiatives [23], but it is perhaps important to reflect that most e-journal preservation initiatives to date have focused on the preservation of the content rather than the interface.

The OAIS model and digital preservation systems

Another important component of a digital preservation solution is the development of organizational models designed to cope with the unique and far-reaching challenges that digital preservation pose. Such organizations will have to be focused on the long-term and adapt to new developments, when necessary. This 'active' approach to preservation is embodied in the definition of digital preservation adopted by the Working Group on Digital Archive Attributes sponsored by the RLG and OCLC Online Computer Library Center. This working group understood digital preservation as "the managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents" [24]. These managed activities depend upon the existence of an organizational entity that can take responsibility for maintaining digital objects. In practice, this means developing some kind of preservation system or repository. In order to be successful, such preservation repositories need to undertake a number of different functions. A start in defining some of these necessary functions has been made by the Reference Model for an Open Archival Information System (OAIS), which has been an international standard since 2003 (ISO 14721:2003) [25].

The OAIS functional model has been used to underpin the development of a number of digital preservation systems. Systems relevant in the e-journal context include the Digital Information Archiving System (DIAS) developed by IBM Netherlands in collaboration with the Koninklijke Bibliotheek (KB) - which forms the basis of both KB's e-Depot and the German KOPAL system - and preservation services like Portico.

Digital preservation strategies

The OAIS Model identifies the main functions that need to be undertaken by preservation services and defines an information model for the objects held by them. However, it does not prescribe the adoption of any particular preservation strategy. The appropriateness of a given strategy depends upon the nature of the object being preserved and the reasons why it is being preserved, i.e. what we have referred to as its significant properties. This means that the choice of a particular strategy, or the exact way that it is implemented, needs careful and expert consideration by repositories.

Thibodeau has developed a spectrum of preservation strategies ranging on a continuum from the preservation of technology to the preservation of objects [26]. In practice, however, most discussion of preservation strategies centres around two main approaches. The most popular of these is migration, in which data objects are continually transformed in order to be usable on new generations of hardware and software. In practice, this approach is often combined with some kind of format standardization undertaken on ingest, a strategy known as 'normalization.' While migration strategies are popular, the fact that objects are subject to almost continuous change means that it is very difficult to ensure that they retain their authenticity [27]. Jeff Rothenberg has argued that migration approaches are labour-intensive, "time-consuming, expensive, error prone, and fraught with the danger of losing or corrupting information" [28]. The second main preservation approach focuses on the emulation of underlying hardware and software environments. Emulation approaches are based on the development of software programs that mimic the behaviour of obsolete hardware and software, so that the original byte-stream can remain usable. Its supporters argue that it is the only reliable way of recreating an object's original functionality or look and feel [29]. Technically speaking, this is far from being a trivial task, but it has been argued that the fact that hardware tends to be well specified at a logical level means that it is an easier task than reengineering application software for new computing environments [30]. The existence of multiple strategies reflects the reality that we do not really know yet which strategies will work best for a given object or preservation objective. They are also not mutually exclusive, meaning that risk can be spread across a number of different strategies. The key thing, whichever strategy (or combination of strategies) is chosen, is to understand that the purpose of any strategy will be to ensure that the significant properties of preserved objects can be retained.

Preservation metadata and packaging models

It has been argued that the key to the successful implementation of all kinds of preservation strategy will be the capture, creation, maintenance and application of appropriate metadata [31]. The type of metadata needed goes far beyond the descriptive metadata traditionally created by libraries, but includes any information that will support the ongoing use and re-use of digital objects. This, so-called, 'preservation metadata' is understood as being all of "the information a repository uses to support the digital preservation process;" specifically, "metadata supporting the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context" [32]. Understood in this way, it is clear that such metadata needs to support an extremely wide range of functions, including recording the contexts and provenance of objects, and documenting repository actions and policies. Over the past decade, there has been a great deal of progress in understanding the metadata requirements of repositories. In this, this OAIS information model has been very influential, not least on the *PREMIS Data Dictionary for Preservation Metadata* published in May 2005.

Central to the OAIS information model is the idea of Information Packages - conceptual objects that securely link objects with their associated metadata. The model defines three different information packages that can be used to support the submission and dissemination of objects as well as for archival storage. The information package concept has informed the development of a number of packaging models for digital objects, including e-journals. In the context of e-journals, much of the focus has been on the development of standardised XML-based packages that can support ingest into preservation repositories.

An early example of this was the XML submission information package (SIP) developed as part of Harvard University Library's E-journal Archiving project, one of a series of seven projects on this general theme funded by the Andrew W. Mellon Foundation. In OAIS terminology, a SIP defines the form of the content that is supplied by a producer - in this case usually a publisher - to an archive or repository. Thus, the Harvard project was primarily focused on the definition of an archival format - in this instance an XML DTD - that could be used for the normalization of source files provided by e-journal publishers [33]. In this particular model, depositing publishers were expected to convert their internal XML or SGML-based source files into this normalised DTD to facilitate transfer into a repository. The Harvard SIP design was based on the XML-based Metadata Encoding and Transmission Standard (METS) and provided a general framework for recording structural relationships between journal issue and item level components, including text and embedded content in other formats (e.g., images or data sets).

The possibility of developing a generic DTD was then taken forward in a project led by the US National Library of Medicine (NLM). The National Center for Biotechnology Information (NCBI), the part of the NLM responsible for biomedical databases, was interested in developing a generic DTD that could be used by the recently launched PubMed Central repository of life sciences literature. Collaborating with XML technology specialists (Inera and Mulberry Technologies), and with the support of the Harvard team, the result of the project was the NLM Archiving and Interchange DTD suite, which has been described as "a set of XML modules that define elements and attributes for describing the textual and graphical content of journal articles as well as some non-article material" [34]. The suite can be used to construct specific DTDs, so NCBI used it to define a Journal Publishing DTD, a 'prescriptive subset' focused on the content submitted by publishers to PubMed Central. Versions of the Archiving and Interchange DTD are also used by a number of small to medium sized publishers, including HighWire Press and the Public Library of Science, and by aggregator services like Ingenta. It also forms a key technical component of the Portico e-journal archiving service, and has been proposed for use by both the Library of Congress and the British Library for migrating electronic journal content to a uniform standard [35].

As this might suggest, XML-based normalization strategies are used by a number of e-journal preservation initiatives. For example, the Portico service has developed an ingest workflow for the capture of publishers source files, producing content and metadata packaged in Portico METS files that can then be ingested into the repository. Portico uses the Archiving and Interchange DTD as a target format for conversion from publishers' DTDs. Evan Owens, Portico's Chief Technology Officer, has

commented that the conversion of publisher DTDs is a complex process, made more difficult by the continued evolution of publishers' formats, meaning that conversion tools need to be frequently updated [36]. E-journal preservation initiatives also attempt to collect as much relevant metadata from publishers as possible. In the Koninklijke Bibliotheek's e-Depot, incoming content and metadata are packaged into 'Publisher Submission Packages.' These are then processed further, with bibliographic descriptions being added to the library's catalogue with other metadata converted into an XML-based format. Members of the e-Depot team have said that, by using the publishers' metadata, "an important labour-intensive task is bypassed" [37]. Owens has noted that Portico's experience is that descriptive metadata is plentiful. He has written that e-journal "articles supplied in marked-up SGML or XML (either full text or headers) normally have all the descriptive metadata clearly identified: author, title, journal, volume, issue, date, etc." [38].

Recent progress in developing e-journal preservation services

As the section on packaging models might suggest, the past five years has seen the continued development of services focused on preserving e-journals and other digital content. In part, this reflects a practical response to digital preservation concerns by a number of national and research libraries, e.g. by the national libraries of Australia and the Netherlands [39]. Other areas of development have evolved out of research activities. Especially important in this regard has been the seven E-Journal Archiving projects funded by the Andrew W. Mellon Foundation. These not only led to a number of co-operative projects with journal publishers and the development of packaging models for the submission of e-journal content, but also led to the detailed investigation of two distinct e-journal repository models, one based on a centralised service (Portico), the other mainly distributed (LOCKSS). To give a flavour of these developments, the following paragraphs will introduce both of these initiatives as well as the Koninklijke Bibliotheek's e-Depot and NCBI's PubMed Central. A recent paper by Anne Kenney provides a brief overview of a number of other e-journal preservation initiatives, including OCLC's Electronic Collections Online, OhioLINK's Electronic Journal Center, and the German KOPAL project [40].

The Portico e-journal archiving service originated in JSTOR's Electronic-Archiving Initiative, a project set-up in 2002 with funding from the Mellon Foundation. In this project, JSTOR spent several years investigating technical requirements and economic models for preserving e-journals and working with publishers on a pilot project, before the Portico service was launched in 2005 with grant support from JSTOR, the Library of Congress, the Mellon Foundation and Ithaka [41]. Central to the service was the development of a sustainable business model. For Portico, this is based on raising revenue from both publishers and libraries to cover ongoing operational costs. Participating publishers provide content (source files) to Portico and are asked for an annual financial contribution based on their total revenues. Libraries also make an annual payment, based on their existing collections expenditure, intended to support the ongoing work of the service. The technical approach is based on the retention of publishers' source files, which are also normalised into the NLM Archiving

and Interchange DTD and packaged into Portico METS files. The service as it normally operates is 'dark' in that it does not routinely provide end-user access. However there are a number of defined 'trigger points' (e.g., if a publisher ceases to operate or a journal title becomes available) that enable access to be provided to participating libraries. The service can also, with the agreement of publishers, be used for providing perpetual access to subscribed content. As of October 2006, nineteen publishers were participating in Portico, including: Elsevier, John Wiley & Sons, Oxford University Press, the American Mathematical Society, and the Institute of Physics Publishing.

At the same time as it funded JSTOR's Electronic-Archiving Initiative, the Mellon Foundation also gave additional funding to Stanford University's LOCKSS (Lots of Copies Keeps Stuff Safe) programme to develop further its distributed approach to the preservation of e-journal content. LOCKSS is a peer-to-peer preservation system based on the existence of multiple low cost persistent caches of e-journal content hosted at the many different institutions licensed to 'own' such content [42]. The system uses the existence of these networked multiple copies to detect and repair damage automatically through voting in "opinion polls." Its supporters have made much of its use of the redundancy inherent in traditional libraries of printed publications. Victoria Reich and David Rosenthal have written: "librarians' defence against irreplaceable loss has always rested on redundancy (one library burns but only one of many copies of a work is destroyed)" [43]. Participating institutions (both libraries and publishers) co-operate through membership of the LOCKSS Alliance, which is a collaborative network based on the open-source software model. LOCKSS takes a different approach from Portico in that it preserves e-journal content in its original form, e.g. as it is harvested from publishers' Web sites. Michael Seadle [44] has argued that by "saving exactly what the reader sees, LOCKSS loses nothing in its archive," while noting the importance of migration as a way of making content available in the future. Various UK higher education institutions are currently experimenting with the system in a pilot programme funded by the Joint Information Systems Committee and the Consortium of Research Libraries, supported by a dedicated LOCKSS Technical Support Service provided by the Digital Curation Centre [45].

As the traditional stewards of the national published output, a number of national libraries have taken a keen interest in the collection and preservation of e-journal content. The institution with, perhaps, the most experience of dealing with e-journal content to date is the National Library of the Netherlands (KB). The KB has had a long-standing interest in digital preservation issues, beginning with its participation in the European Union-funded NEDLIB (Networked European Deposit Library) project in the late 1990s, continuing with experiments on emulation strategies and collaboration with IBM Netherlands on the development of a OAIS-based deposit system for electronic publications. IBM's resulting Digital Information Archiving System (DIAS) formed the basis of the KB's e-Depot system [46]. Following experiments with voluntary deposit arrangements, the KB signed a pioneering agreement with the publisher Elsevier Science in 2002. In this, Elsevier agreed to deposit the content of around 1,300 journals with the KB [47]. Similar agreements have followed with a number of other major journal publishers, including: Springer-Verlag, Blackwell Publishing, Taylor & Francis Group, SAGE Publications, Oxford University Press,

and the open access publisher BioMed Central. While the e-Depot is effectively a 'dark archive,' their agreements mean that the KB does have the right to provide on-site access and document delivery within the Netherlands. It can also provide wider access in the case of publisher or e-journal system failure. Erik Oltmans and Adriaan Lemmen [48] note that the library could provide part of an interim service if co-operating publishers suffered some kind of disaster that made content inaccessible for long periods of time. They add that KB could also provide more permanent access, if the publisher (or its successors) ever stopped making the journals available. The KB's example is gradually being followed by other national and research library-led preservation initiatives. These include the German KOPAL project [49], which is also developing a service based on IBM's DIAS.

An initiative with a slightly different focus is PubMed Central, one of a number of database services provided by the NCBI. PubMed Central was established in 2000, the result of a US National Institutes of Health (NIH) proposal for online services that would provide free access to all biomedical research literature, whether peer-reviewed or not [50]. The controversial nature of the proposed non-peer-reviewed service [51] meant that PubMed Central, when it was eventually established by NCBI, had far more limited aims, namely the provision of a peer-reviewed repository that would provide open-access to the full-text of content published in participating journals. Launching with some extremely high-profile journals (including *Proceedings of the National Academy of Sciences*, *Molecular Biology of the Cell*, and *BMJ*), by the start of 2006 there were over 200 journals participating in the service. PubMed Central allows participating publishers to delay deposit for up to twelve months, but NCBI insist that a journal's participation in PubMed Central is a commitment to open access [52]. Once deposited, PubMed Central is committed to preserving it and maintaining its long-term integrity. In order to facilitate this, it normalises publishers' source files to the NLM Archiving and Interchange DTD. More recently, PubMed Central has become a designated repository for the deposit of research outputs funded by both the NIH and the Wellcome Trust. Also, the Wellcome Trust and a number of other UK biomedical funding bodies have recently awarded a contract to a consortium led by the British Library for the development of a UK PubMed Central service [53].

The wider contexts of scholarly communication

The existence of these ongoing initiatives suggests that there has been considerable progress in developing approaches to the long-term preservation e-journal content. However, the fundamentally interlinked nature of the digital world means that it may no longer be useful to consider journal content in isolation from other forms of scholarly communication. The Internet enables a wide variety of scholarly communication methods, ranging from the formal peer-reviewed paper in an e-journal or conference proceedings, through e-prints stored in online repositories, to the more informal types of communication made possible by technologies like e-mail, wikis and Web logs (blogs). While in the print environment, it was impractical (or unnecessary) to preserve a great deal of this less formal communication [54], the digital world challenges

us to consider anew what particular aspects of scholarly communication need to be preserved. The following paragraphs will briefly explore some of these issues with reference to three main types of content: self-archived papers in e-print repositories, supplementary research data, and Web links.

E-prints

The concept of self-archiving emerged in the 1990s when a growing number of academics and librarians began to promote the idea that the authors of peer-reviewed papers should simply make them available for free by making them available through the Internet. The most frequently cited model of this approach is the subject-based e-print archive first set up by Paul Ginsparg at Los Alamos National Laboratory in 1991 (now hosted by Cornell University and known as ArXiv), a service that initially covered the high-energy physics domain, but which has since expanded to cover other areas of physics, mathematics and computer science. The main focus of interest at the moment is on the development of institution-based repositories. The metadata harvesting standards developed by the Open Archives Initiative (OAI) enable content from multiple institutional repositories to be combined into a single global virtual archive, which Stevan Harnad says makes "all papers searchable and retrievable by everyone for free" [55]. With the practical development of OAI-compliant tools (e.g., repository software like Eprints.org) and the founding of services like PubMed Central, advocacy initiatives like the Public Library of Science [56] and the Budapest Open Access Initiative (BOAI) began to make a high-level case for researchers providing open access (OA) to peer-reviewed research outputs. The BOAI suggested that there were two main ways of doing this: firstly through the deposit of papers in institutional repositories; secondly by publishing in OA journals, whose publishers typically recover costs through combinations of subsidy and author charges. The Directory of Open Access Journals (DOAJ) [57] maintained by Lund University Libraries lists all known OA journals (2,410 in October 2006), including a large number of new titles published by OA publishers like BioMed Central (whose content is already deposited in both PubMed Central and KB's e-Depot) and the Public Library of Science.

OA has become increasingly the focus of policy initiatives led, at least for now, by research funding bodies. For example, the Wellcome Trust - a UK-based charity that funds biomedical research - declared their support of OA principles in 2003 and has since made it a requirement of its grant conditions that funded researchers deposit a copy of research outputs in a designated repository within six months of publication [58]. The designated repository for the time being is PubMed Central, but this will change once the UK PubMed Central service is established. Other funding bodies have begun to follow suit. Following a recommendation from the Appropriations Committee of the US Congress, the NIH has also developed a Public Access Policy that "requests and strongly encourages" funded investigators to make copies of their final, peer-reviewed manuscripts freely available by submitting them, upon acceptance, to PubMed Central [59]. In the UK, a report published in 2004 by the House of Commons Select Committee on Science and Technology recommended that research

councils and other government funding bodies should mandate funded researchers to deposit a copy of published outputs in institutional repositories within a reasonable period of their publication [60]. In response, Research Councils UK consulted on and published a position statement on access to research outputs, the latest version of which (June 2006) enables individual research councils to require funded researchers to deposit outputs in designated repositories [61]. There is also a growing amount of evidence from bibliometric studies that papers freely available online have an impact advantage over non-OA publications [62]. Some self-archiving advocates have used this evidence to argue for the adoption of official university OA self-archiving policies [63]. At the very least, the growing high-level support for OA principles means that e-print repositories look as if they will be a significant part of the scholarly communication system for some time to come.

Proponents of self-archiving emphasise that it is not a replacement for publishing in peer-reviewed journals, but is essentially a supplementary activity focused on enabling OA. For example, Harnad has argued: "authors cannot and should not be expected to stop submitting their research to established high-quality, high-impact journals" [64]. The supplementary nature of e-print repositories means that OA advocates can be hostile to the very idea of long-term preservation principles being applied to the content of e-print repositories. At the very least, Steve Hitchcock, *et al*. argue that "preservation concerns should not be allowed to become a barrier to the deposit of new content" in institutional repositories [65]. That said, however, papers deposited in such repositories are often cited in other research and thus become *de facto* part of the research record. This, and the fact that institutional repositories are seen as potential places for the deposit of other types of institutional content (including research data, learning objects and organisational records), means that preservation concerns cannot be ignored entirely [66]. Clifford Lynch emphasises the preservation role of institutional repositories, arguing that university-based services represent "an organizational commitment to the stewardship of ... digital materials, including long-term preservation where appropriate, as well as organization and access or distribution" [67].

Research data

Similar concerns relate to the long-term curation of research data. Researchers in many branches of science are becoming increasingly dependent on the production and analysis of vast amounts of data, often generated by high-throughput instruments or streamed from sensors and satellites [68]. In addition, as with publications, there is an increasing preoccupation in science policy circles on encouraging open access to publicly funded data. For example, in January 2004, government ministers from all OECD (Organisation for Economic Co-operation and Development) member states endorsed a declaration based on the principle that publicly funded research data should be openly available to the maximum extent possible [69]. Data curation is too large a topic to be dealt with satisfactorily in this chapter, but it *is* relevant because a number of journals now require either the submission of supporting data along with a paper or its deposit in public databases like the Protein Data Bank (PDB) or NCBI's

GenBank. Practical concerns dictate that the institutions that generate data will also have to consider hosting it, at least for the short to medium term, e.g. to comply with the requirements of funding bodies and to defend against accusations of scientific misconduct [70]. Research projects like eBank UK are beginning to experiment with the development of repository models for crystallographic data, but the main focus to date has been on providing ways of publishing data and on enhanced access, rather than on curation [71].

Internet links

A final topic of concern relates to what happens to the Internet references published in journals. A number of studies have demonstrated that links in peer-reviewed journals suffer from severe rates of URL decay (or link rot) [72]. For example, a much-cited 2003 study of links in three major scientific and medical journals (*New England Journal of Medicine*, *JAMA: The Journal of the American Medical Association*, and *Science*) revealed that the percentage of inactive links rose from 3.8 per cent at 3 months to 13 per cent at 27 months after publication [73]. Surveys of URLs in two major computer science journals (*IEEE Computer* and *Communications of the ACM*) and in MEDLINE abstracts have revealed similar trends. The computer science study showed that around 28 per cent of the URLs referenced between 1995 and 1999 were no longer accessible in 2000, rising to 41 per cent in 2002 [74]. The medical study took a slightly different approach, but still showed that in 2003 the overall availability rates of URLs published in MEDLINE abstracts were around 78 per cent [75]. Given these high rates of attrition, it is an open question as to how far this aspect of the integrity of the scientific record can be protected. Proposals include requiring authors to retain printed copies for the short-term and to submit all cited URLs to the Internet Archive (a non-profit organisation that has been collecting Web content since 1996) [76]. Another approach is focused on the development of a new publisher-supported caching service (called WebCite), to which authors would be required to submit URLs before citing them. The system takes a snapshot of the cited page and returns a 'permanent link,' which can then be cited in the published article [77]. It remains to be seen whether either of these approaches will constitute a complete solution to this difficult problem.

Conclusions

This chapter has attempted to sketch out some of the main problems related to the preservation of e-journal content for the long term. The immediate problem relates to the fact that e-access to e-journal content tends to be licensed by libraries rather than owned outright. This problem can be solved to some extent through increased co-operation between libraries and publishers, which needs to focused on the genuine risk of losing e-journal content, e.g. in the case of publisher failure [78]. The LOCKSS initiative and the services provided by Portico, PubMed Central and the Koninklijke Bibliotheek's e-Depot are examples of the kind of joint approaches that

are needed. The longer-term survival of e-journal content will additionally depend on the existence of competent repositories that can take e-journal content from publishers and preserve it through time. While achievable, this is going to be extremely difficult to do. The OAIS model has provided a general framework for the development of preservation services, but it is too early to tell whether existing repositories will be able to fulfil all future requirements. Assuming that they will not, preservation services will constantly have to monitor contexts and technical developments, and respond to changes in appropriate ways. Finally, it is worth remembering that e-journals are just one component of a constantly evolving scholarly communication system and should not be considered in isolation from other developments, e.g. in institutional repositories and data curation. Collaboration and co-operation will be very important in helping to solve these difficult problems. As Brian Lavoie and Lorcan Dempsey have reminded us, digital preservation "is not an isolated process, but instead, one component of a broad aggregation of interconnected services, policies, and stakeholders which together constitute a digital information environment" [79].

Acknowledgements
Work on this chapter was supported in part by grants from the Engineering and Physical Sciences Research Council (GR/T07374/01, Digital Curation Centre: Research) and the European Union's Sixth Framework Programme (FP6-IST-507618: DELOS Network of Excellence on Digital Libraries).

UKOLN is funded by the UK Museums, Libraries and Archives Council (MLA), the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from the JISC, the European Union and other sources. UKOLN also receives support from the University of Bath, where it is based.

References

1. John Ziman, *Public Knowledge: An Essay Concerning the Social Dimension of Science* (Cambridge: Cambridge University Press, 1968), 103.
2. Dale Flecker, "Preserving Digital Periodicals," in *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving* (Washington, D.C.: Council on Library and Information Resources; Library of Congress, 2002), 10-22.
3. Sadie L. Honey, "Preservation of Electronic Scholarly Publishing: An Analysis of Three Approaches." *Portal: Libraries and the Academy* 5 (2005): 59-75, here p. 59.
4. Charles Oppenheim, Clare Greenhalgh, and Fytton Rowland, "The Future of Scholarly Journal Publishing." *Journal of Documentation* 56, no. 4 (2000): 361-98.

5. Dale Flecker, "Preserving Scholarly E-Journals." *D-Lib Magazine* 7, no. 9 (September 2001), http://www.dlib.org/dlib/september01/flecker/09flecker.html (accessed October 9, 2006).
6. Ann Shumelda Okerson, "Buy or Lease? Two Models for Scholarly Information at the End (or the Beginning) of an Era." *Daedalus* 125, no. 4 (1996): 55-76, here p. 68.
7. Model NESLi2 Licence for Journals, http://www.nesli2.ac.uk/model.htm (accessed October 9, 2006).
8. Jennifer Watson, "You Get What You Pay for? Archival Access to Electronic Journals." *Serials Review* 31, no. 3 (2005): 200-319.
9. Honey, "Preservation of Electronic Scholarly Publishing," 60-61.
10. "Urgent Action Needed to Preserve Scholarly Electronic Journals," ed. Donald J. Waters (Washington, D.C.: Digital Library Federation, October 2005), 1, http://www.diglib.org/pubs/waters051015.htm (accessed October 9, 2006).
11. Erik Oltmans, and Adriaan Lemmen, "The e-Depot at the National Library of the Netherlands." *Serials* 19, no. 1 (2006): 61-67.
12. Margaret Hedstrom, "Digital Preservation: A Time Bomb for Digital Libraries." *Computers and the Humanities* 31 (1998): 189-202, here p. 190.
13. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information Commissioned by the Commission on Preservation and Access and the Research Libraries Group* , ed. John Garrett and Donald Waters (Washington, D.C.: Commission on Preservation and Access, 1996), 2.
14. Clifford A. Lynch, "Integrity Issues in Electronic Publishing," in *Scholarly Publishing: The Electronic Frontier*, ed. Robin P. Peek and Gregory B. Newby (Cambridge, Mass.: MIT Press, 1996), 133-45.
15. Alexandre López Borrull, and Charles Oppenheim, "Legal Aspects of the Web." *Annual Review of Information Science and Technology* 38 (2004): 483-548.
16. Abby Smith, *New-Model Scholarship: How Will it Survive?* (Washington, D.C.: Council on Library and Information Resources, 2003), 2.
17. Kenneth Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," in *The State of Digital Preservation: An International Perspective* (Washington, D.C.: Council on Library and Information Resources, 2002), 4-31.
18. Margaret L. Hedstrom and others, "'The Old Version Flickers More': Digital Preservation from the User's Perspective." *American Archivist* 69, no. 1 (2006): 159-87.
19. Margaret Hedstrom, and Christopher A. Lee, "Significant Properties of Digital Objects: Definitions, Applications, Implications," in *Proceedings of the DLM-Forum 2002, Barcelona, 6-8 May 2002* (Luxembourg: Office for Official Publications of the European Communities, 2002), 218-27.
20. For figures from a recent survey of German e-journal publishers undertaken for the German nestor (Network of Expertise in Long-Term Storage of Digital Resources) initiative, see: Gunnar Fuelle, and Tobias Ott, *Langzeiterhaltung digitaler Publikationen: Archivierung elekronischer Zeitschriften (E-Journals)* (Frankfurt am Main: nestor - Kompetenznetzwerk Lang-

zeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland, 2006), 102-11, http://nbn-resolving.de/urn:nbn:de:0008-20051024019 (accessed October 9, 2006).

21. An older study undertaken for the European Union-funded NEDLIB (Networked European Deposit Library) project found that two thirds of the publishers in a non-randomly selected sample generated HTML content 'on-the-fly' from SGML or XML-encoded text. See: Mark Bide & Associates, *Standards for Electronic Publishing: an Overview* (The Hague: Koninklijke Bibliotheek, 2000).
22. Flecker, "Preserving Digital Periodicals," 13.
23. For example, see: Michael Day, "The Long-Term Preservation of Web Content," in *Web archiving*, ed. Julien Masanès (Berlin: Springer, 2006), 177-99.
24. RLG/OCLC Working Group on Digital Archive Attributes, *Trusted Digital Repositories: Attributes and Responsibilities* (Mountain View, Calif: Research Libraries Group, 2002), 3, http://www.rlg.org/legacy/longterm/repositories.pdf (accessed October 9, 2006).
25. *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1 (Washington, D.C.: Consultative Committee for Space Data Systems, 2002), 1-11, http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed October 9, 2006).
26. Thibodeau, "Overview of Technological Approaches."
27. Helen R. Tibbo, "On the Nature and Importance of Archiving in the Digital Age." *Advances in Computers* 57 (2003): 1-67.
28. Jeff Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation* (Washington, D.C.: Council on Library and Information Resources, 1999), 13.
29. *Ibid*., 17.
30. Digital Preservation Testbed, *Emulation: Context and Current Status* (Den Haag: Nationaal Archief, 2003), http://www.digitaleduurzaamheid.nl/bibliotheek/docs/White_paper_emulation_UK.pdf (accessed October 9, 2006).
31. Michael Day, "Preservation Metadata," in *Metadata Applications and Management*, ed. G. E. Gorman and Daniel G. Dorner (London: Facet, 2004), 253-73.
32. PREMIS Working Group, *Data Dictionary for Preservation Metadata* (Dublin, Ohio: OCLC Online Computer Library Center, 2005), ix, http://www.oclc.org/research/projects/pmwg/ (accessed October 9, 2006).
33. Stephen L. Abrams, and Bruce Rosenblum, "XML for e-Journal Archiving." *OCLC Systems & Services* 19, no. 4 (2003): 155-61.
34. National Library of Medicine Archiving and Interchange DTD, http://dtd.nlm.nih.gov/ (accessed October 9, 2006).
35. *Library of Congress, British Library to Support Common Archiving Standard for Electronic Journals* (Washington, D.C.: Library of Congress, April 2006), http://www.loc.gov/today/pr/2006/06-097.html (accessed October 9, 2006).
36. Evan Owens, "Automated Workflow for the Ingest and Preservation of Electronic Journals." Society for Imaging Science and Technology (IS&T) Ar-

chiving Conference, Ottawa, Canada, May 23-26, 2006, http://www.portico.org/
news/Archiving2006-Owens.pdf (accessed October 9, 2006).
37. Erik Oltmans, and Hilde van Wijngaarden, "Digital Preservation in Practice: the *e*-Depot at the Koninklijke Bibliotheek." *VINE: The Journal of Information and Knowledge Management Systems* 34 (2004): 21-26.
38. Owens, "Automated Workflow."
39. Neil Beagrie, *National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and of Related International Activity* (Washington, D.C.: Council on Library and Information Resources, 2003).
40. Anne R. Kenney, "Surveying the E-Journal Preservation Landscape." *ARL Bimonthly Report* 245 (April 2006), http://www.arl.org/newsltr/245/preserv.html (accessed October 9, 2006).
41. Eileen Gifford Fenton, "An Overview of Portico: An Electronic Archiving Service." *Serials Review* 32, no 2 (2006): 81-86.
42. A technical description of the LOCKSS system focusing on its resistance to network attack can be found in: Petros Maniatis and others, "The LOCKSS Peer-to-Peer Digital Preservation System." *ACM Transactions on Computer Systems* 23 (2005): 2-50.
43. Victoria Reich, and David Rosenthal, "Preserving Today's Scientific Record for Tomorrow." *BMJ* 328 (2004): 61-62, here p. 61.
44. Michael Seadle, "A Social Model for Archiving Digital Serials: LOCKSS." *Serials Review* 32 (2006): 73-77.
45. Helen Hockx-Yu, "Establishing a UK LOCKSS Pilot Programme." *Serials* 19 (2006): 47-51.
46. Oltmans and van Wijngaarden, "Digital Preservation in Practice," 21-23.
47. Oltmans and Lemmen, "The e-Depot at the National Library of the Netherlands," 61-67.
48. *Ibid*., 64.
49. KOPAL, http://kopal.langzeitarchivierung.de/ (accessed October 9, 2006).
50. The controversy surrounding the origins of PubMed Central are investigated in: Addeane S. Caelleigh, "PubMed Central and the New Publishing Landscape: Shifts and Tradeoffs." *Academic Medicine* 75, no. 1 (2000): 4-10; and: Rob Kling, Lisa B. Spector, and Joanna Fortuna, "The Real Stakes of Virtual Publishing: The Transformation of E-Biomed into PubMed Central." *Journal of the American Society for Information Science and Technology* 55 (2004): 127-48.
51. For example, see: Arnold S. Relman, "The NIH 'E-Biomed' Proposal - A Potential Threat to the Evaluation and Orderly Dissemination of New Clinical Studies." *New England Journal of Medicine* 340, no. 23 (10 June 1999): 1828-29.
52. David L. Wheeler and others, "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 34 (2006): D173-80.
53. UK PubMed Central project, http://www.wellcome.ac.uk/doc_wtd015366.html (accessed October 9, 2006).

54. John Ziman, *Real Science: What it Is, and What it Means* (Cambridge: Cambridge University Press, 2000), 265.
55. Stevan Harnad, "The Self-Archiving Initiative." *Nature* 410 (26 April 2001): 1024-25, here p. 1025.
56. Richard J. Roberts and others, "Building a 'GenBank' of the Published Literature." *Science* 291 (23 March 2001): 2318-19.
57. Directory of Open Access Journals, http://www.doaj.org/ (accessed October 9, 2006).
58. Robert Terry, "Funding the Way to Open Access." *PLoS Biology* 3, no. 3 (2005): e97, 364-66.
59. *Implementation of Policy on Enhancing Public Access to Archives Publications Resulting from NIH-Funded Research* (Bethesda, Md.: National Institutes of Health, April 2005), http://grants.nih.gov/grants/guide/notice-files/NOT-OD-05-045.html (accessed October 9, 2006)
60. House of Commons, Select Committee on Science and Technology, *Scientific Publications: Free for All?* (London: The Stationery Office, 2004).
61. Research Councils UK, *Updated Position Statement on Access to Research Outputs* (London: RCUK, June 2006), http://www.rcuk.ac.uk/access/2006statement.pdf (accessed October 9, 2006).
62. For example in: Steve Lawrence, "Free Online Availability Substantially Increases a Paper's Impact." *Nature* 411 (31 May 2001): 521; and: Gunter Eysenbach, "Citation Advantage of Open Access Articles." *PLoS Biology* 4, no 5 (2006): e157, 692-98.
63. Stevan Harnad and others, "The Access/Impact Problem and the Green and Gold Roads to Open Access." *Serials Review* 30 (2004): 310-14.
64. Harnad, "Self-Archiving Initiative," 1025.
65. Steve Hitchcock and others, "Preservation for Institutional Repositories: Practical and Invisible," Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data (PV 2005), Edinburgh, UK, November 21-23, 2005, http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/033.pdf (accessed October 9, 2006).
66. Raym Crow, *The Case for Institutional Repositories: a SPARC Position Paper* (Washington, D.C.: Scholarly Publishing & Academic Resources Coalition, 2002), http://www.arl.org/sparc/IR/ir.html (accessed October 9, 2006).
67. Clifford A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." *ARL Bimonthly Report* 226 (February 2003), http://www.arl.org/newsltr/226/ir.html (accessed October 9, 2006).
68. For example, see: Tony Hey and Anne Trefethen, "The Data Deluge: An E-Science Perspective," in *Grid Computing, Making the Global Infrastructure a Reality*, ed. Fran Berman, Geoffrey Fox, and Tony Hey (Chichester: Wiley, 2003), 809-24.
69. Peter Arzberger and others, "An International Framework to Promote Access to Data." *Science* 303 (19 March 2004): 1777-78.
70. Philip Campbell, "Electronic Futures in Scientific Communication and Outreach." *Journal of Molecular Biology* 319 (2002): 963-67.

71. Simon J. Coles and others, "An e-Science Environment for Service Crystallography: From Submission to Dissemination." *Journal of Chemical Information and Modeling* 46, no. 3 (2006): 1006-16.
72. Steve Lawrence and others, "Persistence of Web References in Scientific Research." *Computer* 34, no. 2 (February 2001): 26-31.
73. Robert P. Dellavalle and others, "Going, Going, Gone: Lost Internet References." *Science* 302 (31 October 2003): 787-88.
74. Diomidis Spinellis, "The Decay and Failures of Web References." *Communications of the ACM* 46, no. 1 (January 2003): 71-77.
75. Jonathan D. Wren, "404 Not Found: the Stability and Persistence of URLs Published in MEDLINE." *Bioinformatics* 20, no. 5 (2004): 668-72.
76. Kathryn R. Johnson and others, "Addressing Internet Reference Loss." *The Lancet* 363 (21 February 2004): 660-61.
77. Gunther Eysenbach, and Mathieu Trudel, "Going, Going, Still There: Using the WebCite Service to Permanently Archive Cited Web Pages." *Journal of Medical Internet Research* 7, no. 5 (2005): e60, http://www.jmir.org/2005/5/e60/ (accessed October 9, 2006).
78. "Urgent Action Needed to Preserve Scholarly Electronic Journals," 2.
79. Brian Lavoie and Lorcan Dempsey, "Thirteen Ways of Looking at ... Digital Preservation." *D-Lib Magazine* 10, no. 7/8 (July/August 2004), http://www.dlib.org/dlib/july04/lavoie/07lavoie.html (accessed October 9, 2006).