

Archiving in the Digital Era

John Frow

[Go back](#) to the AHRB Centre's publication section.

I'm neither a lawyer nor a librarian; what I'm doing in this paper is putting together pieces of knowledge from fields in which I have no expertise, so I make no promise that my knowledge is entirely accurate or entirely up-to-date.

My interest is in the commodification of culture, and more particularly of information; and thus, and only so far, is it an interest in the institution of intellectual property.

It's because of this, though, that I have an interest in the library and other archives as something like a model of open access to information, of a kind which is in many ways threatened by the increasing commercialisation of knowledge.

This is the subject of a conference on the digitalisation of the archive that I'm running next month; this is a paper not about what I know but about what I hope to find out at the conference.

Let me start by asking a question around the answers to which my talk will be shaped: the question, What is an archive?

First answer: an archive is a repository of copies of works ie, of copyrighted and once-copyrighted materials, which have been acquired by deposit or gift or purchase, and which are available for public use (consultation or borrowing) under specific forms of legal exemption from the limitations of copyright:

In the areas of the world governed by the WIPO convention there are three main areas of exemption. The first is the First-sale doctrine, which stipulates that once the work has been legitimately copied and published, the author's rights in relation to that copy are exhausted. The second is the provisions for fair use / fair dealing, primarily for educational and scholarly purposes. And the third is the libraries exemption (section 108 of the US Copyright Act 1976; amended in DMCA, 1998; in UK, sections 37-44 of the Copyright, Designs and Patents Act 1988), allowing copying by the so-called prescribed libraries under certain very strictly circumscribed circumstances: in simple terms, what is allowed is the making of single copies of articles or short extracts from books for research or private study, and of replacement copies for preservation purposes or where an item has been lost, but only where the item is no longer available on the market.

A second and broader answer to my question would say that an archive is a repository of information, which takes the form of copies of works which are regulated by a particular intellectual property regime.

Information has in most societies been culturally framed as an inappropriate object of private and exclusive ownership, although in the societies that we think of as constituting Western 'modernity' this ethos has coexisted with the partial monopolies granted by copyright law and other forms of intellectual property rights. There has, however, been increasing commercial pressure for the privatisation of these categories and their removal from the commons.

The characteristic structure of information is that of gift exchange without monetary recompense but in contexts of calculation and strategic manoeuvre. The barter of information, and the estimation of who knows what, condition the forms of its reciprocity but can never endow it with real scarcity, since its most important quality is its inexhaustible reproducibility: if I tell something to you I still 'possess' it myself, and so on indefinitely. Although the category of intellectual property is grounded

in an extension of the concept of real property, the fact is that, unlike land or material goods, information is not consumed by use.

Thomas Jefferson put this much more eloquently than I can in a letter of 1813:

" If nature has made any one thing less susceptible than all others of exclusive property, it is the action of the thinking power called an idea, which an individual may exclusively possess as long as he keeps it to himself; but the moment it is divulged, it forces itself into the possession of everyone, and the receiver cannot dispossess himself of it. Its peculiar character, too, is that no one possesses the less, because every other possesses the whole of it. He who receives an idea from me, receives instruction himself without lessening mine; as he who lights his taper at mine, receives light without darkening me. That ideas should freely spread from one to another over the globe, for the moral and mutual instruction of man, and improvement of his condition, seems to have been peculiarly and benevolently designed by nature, when she made them, like fire, expansible over all space, without lessening their density at any point, and like the air in which we breathe, move, and have our physical being, incapable of confinement or exclusive appropriation. Inventions then cannot, in nature, be a subject of property."

This indeterminacy of the positioning of information (its ability to exist in many places simultaneously) gives rise to a more general indeterminacy inherent in the fact that information is structured as an open system with multiple users; its 'value' can be assessed only retrospectively in relation to its contexts of use. Two obstacles to commodification develop immediately from this radical indeterminacy. The first is the problem of defining and enforcing exclusive property rights (which are the precondition for capital investment) in something both intangible and diffuse. The simple solution to this problem is to treat information as a secret. The more complex solution, and the one that has been worked out in great detail by Western law over the last three centuries, is to restrict access to and use of information without necessarily restricting possession of it; thus copyright law restricts only the making of unauthorised copies, patent and trademark law restrict commercial exploitation, and so on.

The second obstacle is the problem of attaching exchange value to an entity which has an almost limitless use value: that is, of making an abundant good scarce. The uncertainty that flows from the indeterminacy of uses (the unpredictability of the 'take' of any information product) entails considerable risk for capital investment. At the same time, the relatively high costs of initial production and the relatively low costs of subsequent copying of information goods make predictability imperative. The problem of the minimisation of risk can be solved in part by generating 'a constant stream of unique (if often similar) products with severely limited life spans', and more broadly by a combination of control of access and regulation of demand through the mechanisms of the law. Nicholas Garnham has detailed five strategies elaborated by the culture and information industries to this end. They are: (1) the production of scarcity by controlling the right to copy; (2) the control of distribution channels; (3) the building-in of obsolescence through the manipulation of time (for example by the creation of rapidly decaying information in journalism); (4) the sale of audiences to advertisers rather than, directly, of cultural and informational goods to consumers; (5) the socialisation of production costs by means of state patronage and subsidy. To these we should add the massive work of classification and certification effected by the institution of authorship, which remains the single most important channel for the creation of textual desire and the minimisation of market uncertainty.

For all libraries except deposit libraries, such as the National Library of Scotland, a condition of their successful existence is that they negotiate a path between the costly scarcity induced by these strategies and their role of making information broadly accessible to a public.

A third kind of answer to my question about what an archive is, then, would look at the way libraries and other archives work both as concrete institutions and as a model of one set of social relations for

the circulation of knowledge: the so-called library model of knowledge management, in which a hybrid mode of free or nearly free lending threads a middle way between the economy of the commodity and the economy of the gift.

A library in this sense is a collection of informational materials, traditionally but not necessarily printed matter, which have typically been bought in the market but which, in most public library systems, do not circulate as commodities. But neither do these materials circulate as gifts; they are, rather - to use Marcel Mauss's term - prestations, 'gifts' that return without conferring any rights of ownership or permanent use. At the same time, loaned library materials create no personal ties of obligation and lack the coerciveness of the forms of prestation that Mauss describes for traditional gift economies. In this sense, they partake of the impersonality and the abstractness of the commodity form; unlike commodities, however, they have also been largely free of the forms of coercion (the constraints on access and use) that tend to flow from the price mechanism. While the 'library model' thus tends to collapse rather than to dichotomise the categories of gift and commodity, it does nevertheless represent a genuine alternative to the privatisation of the commons in information. This model is, however, under threat.

'The current phase of capitalist development', says Gareth Locksley, 'is one characterised by the elevation of information and its associated technology into the first division of key resources and commodities. Information is a new form of capital', and as such it undergoes a change of form: rather than being deposited primarily in an interlocking ensemble of open, 'library' systems with minimal entry requirements, it is increasingly managed within a system of private ownership where access is regulated by the payment of rent.

Public libraries as we know them came into being as part of that massive expansion of state institutions in mid-nineteenth century Europe and North America that also produced the public schooling system, post offices, railways, and public hospitals, and which set an ethos of public service against the monopolistic tendencies of the uncontrolled market. Their present existence is framed by a tension between that expanded model of the state and its role in the provision of free (that is, subsidised) public services and a more restrictive view of the state which seeks to open the provision of information to market forces. To put it crudely, a model centred on the informing of citizens has been replaced, at least in part, by a model of choices made by consumers. The causes of this shift are many and complex, but a major one in the case of the public library system has been that change in the status of information itself, from being 'economically valueless, mainly government produced and largely public, to being value-added, commercially sensitive and high cost'.

In an American study, Herbert and Anita Schiller identify a 1982 U.S. Government report as a turning point in the progressive weakening of the 'library model'. Announcing an end to the principle of cooperation between the public and private information sectors, the report represents 'the private industry's challenge to the right of the public sector (government, libraries, universities, etc.) to engage in any activities the industry regards as its own province' - that is, any activities which might have a commercial potential. The screws are thenceforth on the public library system (perhaps the most genuinely popular of all cultural institutions) not only to implement various local forms of commercial practice - 'charging users for information, relying on private vendors for databases, contracting out functions to private firms, and so on' - but more generally to relinquish its primary role in the provision of information.

The focus of my talk today is on the tension between this impetus to make information scarce and expensive, and the electronic revolution - a revolution in the technologies of proliferation and dissemination - which has generated the possibility of a vast increase in the accessibility and visibility of information, and in particular of a transformation of the archive into something like a model of the public domain in which knowledge circulates freely: a model, let me add, which corresponds closely to that Enlightenment ideal of open knowledge systems which underlies the Western scientific ethos.

I'm concerned here not with the totality of the electronic revolution but with its specific effects on archives; in particular, I'm not directly concerned with the internet except as it concerns the storage and retrieval of research materials. Two main areas are of direct relevance here: the effect of intellectual property law on the workings of archives of different kinds; and the role of intellectual property in the current crisis in scholarly and scientific publication.

At the centre of the electronic revolution is the technology of digitalisation: that is, the transformation of analogue print or visual or auditory materials into a common machine-readable form.

The digital in this sense is a 'copy' of an original, produced by the keying in or scanning of documents, by downloading or transferral from another database, or by the networking of material. This means in turn that the distinction between 'reading' and 'copying' which governs the use of print becomes far more difficult to define and enforce in the digital domain, and it means that one of the salient effects of digitalisation is the creation of new forms of property relation: its proliferation of copies at virtually zero cost tendentially undermines the tight control of copying that is at the centre of Western intellectual property regimes. Despite these consequences, however, the institutions of intellectual property continue to work within a framework designed for print, and for the figure of the individual author who uses publishers as mere intermediaries.

Amongst other things, this means that many copyright exemptions for print have not been extended to digital copying. Libraries' limited rights to copy print materials for their users are not available for electronic copying; so any use of copyright digital information must be licensed by the rightholder. In the US, Digital copying is allowed for preservation (ie, for 'archival purposes', effectively the replacement of an owned original copy), but this archiving right doesn't include the right to convert print and other non-digital media to electronic form for storage and retrieval purposes: the DMCA amendment to section 108 allows libraries to make 3 copies, but not for use outwith the library; that is, libraries have no right to lend digital copies of works, which means the first-sale doctrine doesn't apply to them.

Similarly, the European Commission Directive on the harmonisation of copyright makes it possible, at least in some countries, for libraries to digitalise materials in their collection but not to communicate it to the public without explicit permission - something which is very difficult in the case of older materials or material published outside mainstream sources, and especially given the fact that in the current circumstances of commercial uncertainty publishers have often been unwilling to give any permissions for digital reproduction.

Finally, there is no US or EU right to make back-up copies of audio and video tape recordings: libraries can only replace damaged copies, and only if there is no market alternative.

The general effect of this is a de facto prohibition on electronic browsing (since only the copyrightholder may 'communicate to the public'); there might indeed well be cases where the paper form of a work could legally be made available to the public but not the electronic form; digitalisation takes place at two levels: that of metadata and that of data; the first, metadata, would include the making of catalogues, the construction of search engines, the downloading of content from search engines, abstracting, and the creation of bibliographies. The level of data refers to the digitalisation of existing content, or the creation of new digital content, as part of the archival function of libraries.

Carol Henderson, the President of the American Libraries Association, recently wrote about this: 'Libraries ... recognize that a key societal function of libraries - the archival function - is at risk because electronic information is so seldom actually available for purchase and permanent retention or preservation. Libraries play this archival role because history has shown that it is not economically viable for profit-based businesses to do so. The disappearance of much electronic

information after a very short period of time, the fragility of digital bits, and the short life of hardware and software suggest that this role of libraries will be more needed than ever before, but harder and harder for libraries to accomplish.'

There are a number of issues in play here:

1. The question of archiving rights to electronic journals: university libraries often need to pay several times: for the journal subscriptions; for the photocopying licence to enable them to make copies for teaching purposes; for clearance to include the same material in study packs. Yet, after these various levels of licensing, they still don't necessarily have a collection of archived back issues
2. Increasingly, libraries are purchasing or leasing digitalised materials which are based on collections held in other libraries: it's the intermediaries that are profiting from this exploitation of the value of special collections
3. As resource archives, libraries increasingly make use of digitalised teaching resources, especially multimedia resources. A key problem here is that a CD-ROM, for example, may incorporate the rights of multiple authors, including the developer of the underlying software; and, after the Copyright, Designs and Patents Act of 1988, any author whose work is incorporated in another work become rightholders: for example, in the case of a sound recording of a musical performance, the owners of the copyright in the 'underlying works' represented by the music and the lyrics will have to give their consent before the recording can be lent to the public by a public library
4. Distance education raises particular problems to do with the provision of digital materials, since these require the making of copies and, in addition, may render the provider liable as an online service provider.
5. One of the most important activities of libraries is document delivery. Under the constraints of current intellectual property law, this can only be construed as a commercial activity, licensed by rightholders and thus dependent on the granting of permissions.
6. The lending of software (at least where the software is less than 50 years old) is prohibited in most contemporary intellectual property regimes even though it's clearly a 'literary work' in sense of the relevant Acts. There is no exemption for fair dealing in the case of software, and there are problems in lending books with software attached.
7. Finally, there is a whole different set of problems associated with image archives and with the collection and holding of material objects, where different sorts of intellectual property rights obtain and where, crucially, the dimensions of work and copy coincide.

The most general level of the transformation of the function of the archive concerns the evolution of new conditions of access, or restriction of access, for digital material.

Over 50% of income collected by the Authors' Licensing and Collecting Society on behalf of authors derives from photocopying royalties from the print medium; significant moves are now under way to extract royalties from digital reproduction: from the licensing of digital archive copies, of works made available to disabled people, of the first digitalisation of a work, and of value-added services such as the provision of course packs.

In one sense, the use of a license mechanism is a technical question about the facilitation of permitted exemptions to the limitations of copyright; but in another sense it has economic consequences which are potentially crippling for public libraries and other archives.

This brings me to my second area of concern: the transformation of scholarly and scientific

publishing In many domains, especially the sciences, print publication is being phased out, to be replaced not just by electronic journals but by powerful electronic interfaces. Publishers like Reed-Elsevier (Science Direct) and OCLC (First Search) deliver content - both metadata and data - from vast databases (First Search, eg, references over 5.9 million online full text articles from over 3,500 electronic journals); libraries are licensed intermediaries for these databases, but in one sense act merely as brokers to squeeze better deals from publishers. One consequence of this is that [Nature 397 (Jan 1999)] the ability to click from abstract or citation to the full text of an article is prompting a shift in the way that journals are used. Scientists often care less about the journal title than the ability to track down quickly the full text of articles relevant to their interests. Increasingly, users view titles as merely part of hyperlinked 'content databases' made up of constellations of journal titles.

In high-energy physics the Los Alamos e-print archives have become the primary mode of publication: here, articles are published as pre-prints, with peer review coming after publication; a similar model is currently being developed for the biomedical sciences. These moves are of course in part a response in part to the 'serials crisis'

By chance, a circular arrived on my computer on Tuesday from the Consortium of University Research Libraries which sends out a cry for help in relation to the science journals crisis: 'The advent of electronic publishing, combined with the concentration of research publications in Science, Technology and Medicine (STM) in the hands of a small number of publishers, has led to a situation whereby the product of scholarly endeavour is increasingly owned and controlled by commercial publishers, whose priorities are driven by shareholder interests.' The circular goes on to warn of a significant inflation in periodical prices, amounting to a 291% increase, during the period between 1986 and 2000.

Science publishing is now dominated by three huge conglomerates: Elsevier, Thomson, and Bertelsmann. Elsevier controls something like 1500 journals, incl. key journals and databases in medical science, biology, and business; it had profits of £230 million on sales of £571 million in 1997 from its scientific activities alone. Its title Brain Research, to take one example, doubled in cost between 1992 and 1996 to US\$15,000 annually. The US Association of Research Libraries calculates that its 114 member libraries spent 142% more on journals in that year than 10 years before but ordered 6% fewer titles.

The effect on the budgets of research libraries has thus been catastrophic: they are massively distorted towards the costs of scientific journals and database licenses, and there have thus been extensive cut-backs in monograph purchases - which in turn hurts scholarly publishers, which in turn makes it much harder for young scholars to publish, and so on down the line.

The situation is untenable, and it has generated a number of attempts to circumvent it by means of alternative forms of organisation of publishing in 1995 Stanford created High Wire Press to return scholarly publishing to non-profit organisations SPARC, the Scholarly Publishing and Academic Resources Coalition, set up in 1997 by the US Association of Research Libraries, is underwriting the launch of titles directly competing with expensive titles, with its members committed to buying each of them Its electronic chemistry journal, PhysChemComm, for example, sells for \$353 and competes directly with Elsevier's Chemical Physics Letters at \$8,000 The Public Library of Science has similar aims: academics who have signed up to this initiative have promised to publish in, edit or review for, and personally subscribe to, only those scholarly and scientific journals that have agreed to grant unrestricted free distribution rights to any and all original research reports that they have published, through PubMed Central and similar online public resources, within 6 months of their initial publication date.

The other document that hit my desk this week was the Budapest Open Access Initiative, sponsored by George Soros's Open Society Institute. Its preamble says: 'An old tradition and a new technology

have converged to make possible an unprecedented public good. The old tradition is the willingness of scientists and scholars to publish the fruits of their research in scholarly journals without payment, for the sake of inquiry and knowledge. The new technology is the internet. The public good they make possible is the world-wide electronic distribution of the peer-reviewed journal literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds. Removing access barriers to this literature will accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich, make this literature as useful as it can be, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge.' The initiative goes on to call for an extension of such access to all of the scholarly and scientific literature, through two mechanisms: self-archiving, and the development of alternative journals which don't charge subscription or access fees and don't use copyright to restrict use of the materials they publish.

What is at issue here is the lopsided market situation in which the costs of generating intellectual property are met by, mostly, public institutions, and the intellectual property rights are then donated to publishers who re-sell it to university libraries at a high cost John Sutherland, writing in the London Review of Books in 1999, put it this way:

" It's a sweet deal for the publisher, who pays none of the costs of originating his material. Those costs, which can run into millions of dollars and years of salaried time, are picked up by the authors' institutions or by grant-awarding bodies. Authors are paid nothing for the publication of their work. Nor do journals normally pay for the confidential peer reviews which guide their selection. The publisher thus gets an excellent product gratis, and all he has to do is package it. And, sweetest of all, his running costs and overheads are covered by subscriptions, the level of which he himself sets. Effectively, this means that he can make universities pay through the nose for something that the universities have paid to produce in the first place. "

Any final solution to this problem rests with the universities making use of the market leverage given them by the intellectual property rights that they themselves generate.

Caltech, to take one example, has proposed that in the short term it wants all its faculty to agree that they will publish in journals only on the basis that they and Caltech, as joint copyright-holders, lease the material to the publisher for a limited period (reversion after two years has been proposed).

More generally: it seems to me likely that any solution to the crisis in the cost of scientific publication will come only when the world's major research universities band together and use their market power to overcome these market distortions, as well as using their political leverage to restrain the claims that publishers are making to extensive intellectual property rights (and this then returns us to the fundamental question of the purpose of copyright law: the question of whether it is a mechanism for creating and defending property rights, or whether the accent should be on the limited nature of copyright, its function as a necessary evil serving the larger end of a flourishing public domain).

In the Preface to A Contribution to the Critique of Political Economy Marx wrote: 'At a certain stage of their development, the material productive forces of society come in conflict with the existing relations of production, or - what is but a legal expression for the same thing- with the property relations within which they have been at work hitherto. From forms of development of the productive forces these relations turn into their fetters.'

This seems to me precisely the situation here. It's not the institution of intellectual property as such that's the problem, but the form that it takes - the particular, historical balance between the limitations it imposes and the exemptions to them, the balance between rights and the qualifications of rights.

The struggle to redress this balance is one that goes to the core of our work as scholars with a basic commitment to the scientific ethos of the free circulation of knowledge.