Date : 16/07/2008

Electronic publication: Problems of Archiving and Access to Archived Information: legal deposit, data protection and related topics

Harald von Hielmcrone
State and University Library
Aarhus, Denmark

Meeting: **87 Copyright and other Legal Matters with FAIFE**
Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

INTRODUCTION

Electronic publishing presents libraries and archives with new and serious problems when it comes to preserving literary heritage and securing access to information. These problems relate both to the media used and to the legislation regulating the use of the media.

Since the invention of writing physical media have been used for preserving texts, and since the invention of printing paper has been the preferred medium, for both printed and written texts. The fixation of the text on a physical medium like a piece of paper or a printed book has obvious advantages for libraries and archivists. There are physical limits as to what you can do with them.

One of the limits is that a published book will forever remain available to the public unless all existing copies perish, and this is not likely to happen in countries with legal deposit of printed works. This is due to the simple fact that copies of the book have been widely distributed – sold to whoever wanted to buy them and so, once sold, the book is available, and in the public domain.

This fact has been reflected in copyright law. The author has the exclusive right to decide whether a literary work he has produced is to be published or not. But when a book has once been published, the author's legal right to control the distribution has been "consummated" or "exhausted".

When it comes to electronic publishing all this is different.

A work in digital form is not fixed in the same way as a printed text is. It may be moved via network to any desired place within seconds and at no cost or effort at all, and you may without any effort produce as many copies as you like. The fixation to the medium has become irrelevant. The work can be regarded as a free floating and truly immaterial object.

Instead of publishing a book, you may now choose to publish the same contents electronically in a database and make them accessible via Internet. If you regret it, you may at any time unload it from the database. You may also choose to give access only to certain people and not to others. You remain at any time in control of whom – if anybody - shall be given access to the work.

The difference between what you can do with a physical object and what you can do with an immaterial object is reflected in the copyright rules concerning author's rights. According to the WIPO Copyright Treaty, authors right to control the communication to the public of their works is not consummated or exhausted. This means that, in principle, every time you want to access a work that is published electronically in a database you need the author's permission.

Many libraries have made the move from "collections" to "connections". Instead of acquiring a copy of an electronic book or journal and physically installing it on the library server, the library's subscription allows its users to accesses the material via an internet address. This has obvious advantages, but it does remove access control from the hands of the library and places libraries and their patrons at the mercy of suppliers and authors.

It is reasonable to expect that if there is a commercial interest in a work it will remain available. But use of a work may become so infrequent, that it no longer covers its costs. Many libraries try to overcome this issue by obliging suppliers to guarantee "eternal access". Such guarantees are worthless, however. In the first place, the supplier may not be able to fulfil these obligations, for example, if the supplier goes out of business. Secondly, the author may enforce the right of communication to the public and withdraw the work. This may happen, for example, when an author regards an earlier work as a youthful aberration whose contents or quality does not meet the author's present standards. If the work is published in print form, there is nothing the author can do. But if it is published electronically in a database, the work may simply be removed. The consequences for historical research are obvious.

The development of digital technology has also brought about the development of a completely new legal area: data protection. Of course, there has since Roman times been legislation against slander, and archives are governed by laws and regulations in order to protect the privacy of persons mentioned in archived material. But data protection is something different. It does not only concern defamations or information about intimate details of a people's private lives. Data protection applies to all information that can be related to a person, thus including completely ordinary and publicly available information.

The reason behind the development of data protection legislation is the fact that it is possible by computerised means to collect available information related to individual persons, and by combining this to make such detailed pictures of person's present life and

history that many feel uncomfortable, and protest a violation of privacy. The development of the internet and extremely efficient search machines has aggravated the situation.

The main principle of data protection is that personal data may only be processed with the consent of the person they relate to, the data subject. There are exceptions to this requirement, but this is where you start. And when it comes to personal data, revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and data concerning health or sex life, the possibilities for access are very limited. Making such data available to the general public for undefined purposes is out of the question – at least in Europe.

These examples highlight what is at stake in when we move from printed to electronic publication. We face the questions whether:

- the general public shall have freedom of access to published information;
- libraries and archives shall have the right to preserve the cultural heritage of electronic publications; and
- historical research on the basis of original and undistorted sources can be guaranteed.

This paper describe the challenges this legislation present libraries and archives with in their efforts to preserve and secure access to archived electronic information.

LEGAL DEPOSIT AND ELECTRONIC PUBLICATIONS

There are different ways to secure the literary heritage of a country. Since the Enlightenment many countries have a system of legal or voluntary deposit of published books and journals and other kinds of printed material like pamphlets, posters etc. In Demark legal deposit was introduced in 1697. The initiative came from the royal Librarian, who wanted copies for the library and for use in exchange for foreign books from other libraries. Printers had to deliver five copies of everything they printed. Already 30 years later the number of copies to be delivered was reduced to serve only domestic needs. The system continued with small adjustments for 300 years until two major revisions, in 1997 and again in 2004. The main reasons for the revisions were the development of electronic publishing and the internet. The revisions were made in two steps, and it may be instructive to look at these separately.

The 1997 revision

Until 1997 the Legal Deposit Act had only included printed material. In the latest version before the revision printers had to deliver 2 copies of everything they printed to the national libraries and on request one copy to the Library of Natural Science and Medicine. The main objective of the 1997 revision was to have non-printed material included in legal deposit. The law then was reformulated so that two copies of all *published* works had to be delivered to the national libraries. This definition included all works published on physical

media like, recorded music on CD, published film on whs tapes or DVD and text, and multimedia on CD-ROM.

The definition, however, also included works that were published in databases, i.e. uploaded in databases and made available to the public via the internet. At this time the internet was a relatively new technology for use by the general public, and it was not obvious how the concept of legal deposit could best be applied to the contents made available here. Especially many internet sites and homepages were dynamic, being frequently changed or updated. It was difficult to say, what was to be "deposited" and equally difficult to say how this should eventually be done.

It was decided to distinguish between static and dynamic works, i.e. works that were finished and not meant to be changed, e.g. journal articles, reports and e-books, and dynamic works that were designed to be constantly changing, e.g. newspaper homepages and databases. Publishers of static works like journal articles, reports and e-books were by the new law obliged to report the uploading of such works, i.e. the publishing of them, to the Royal Library, and give the library access to download the material.

A system was developed to manage this process and it functioned rather smoothly. However, only a small proportion of all the static works published were reported by the publishers, and as a result it was not possible to reach the level of comprehensiveness that you would require from a legal deposit system.[1]

The 2004 revision

Whereas it was relatively easy to identify printers and producers of music, film and multimedia, and to enforce the delivery of the required copies, internet publishing was completely decentralised. There were in 1994 more than 500.000 Danish domain names (.dk). Which of all these sites, all these "publishers", had uploaded static works to be reported to the legal deposit system of the Royal Library and downloaded? Impossible to say and impossible to manage. Other solutions had to be looked for.

By this time the harvesting technology had developed to a degree that it became feasible to imagine the harvesting of the Danish part of the internet. The State and University Library of Aarhus conducted in 2003 some tests which demonstrated the possibility and also that it was possible at a reasonable price.

In the 2004 version of the Legal Deposit Act the sections concerning the reporting of publishing electronic publications in databases were substituted with a new chapter on harvesting the Danish part of the Internet. The principles for harvesting were:

- cross section,
- selected sites,
- selected events.

[1] From 1998 to 2002 only 14,301 units were reported to the system.

Cross section harvesting means that the whole Danish part of the internet is harvested four times pr year. The harvester searches primarily for .dk domains, but there are also other criteria for determining whether a site is Danish.

Selected sites means that some 80 sites are selected according to certain criteria, and these sites are harvested completely, i.e. all changes. The most prominent, of this category and by far the largest are the sites of Danish newspapers and broadcasting corporations.

Selected events mean that if events of some political or historical importance take place, such sites are harvested completely for a period. The Cartoon Crisis of 2006 is perhaps the most prominent example of an event of great political and historical importance for Denmark.

The harvested material is collected in a database, and the plan is that it shall be indexed and searched in the same way as the living internet. However, the tools for indexing and searching have not been developed yet, so for the time being material may only be accessed via the domain name.

LEGAL ASPECTS OF HARVESTING THE INTERNET

Copyright

Internet sites are, of course, protected by copyright. The individual works e.g. books, articles, pieces of music etc. made available on a site are created by authors who have the right to authorise – or refuse to authorise – the copying and the communication to the public of their works. There may be neighbouring rights, e.g. those of performing artists of music and film whose permissions are necessary, and even the producer of the database has the right to authorise – or refuse to authorise – any substantial extraction of the contents of the database.

With so many internet sites and so many rights holders involved in every internet site it is impossible to get permission to harvest, i.e. copy, the contents of every internet site, and therefore any harvesting of the content of internet sites requires a legal authorisation that overrides the protection offered by the copyright legislation.

Data Protection

Many, maybe even most internet sites contain personal data, i.e. data that relate to living identifiable persons. The legal regulation of the processing of personal data differs considerably. In Europe data protection is quite strict, and covers all member states of the European Union.

The definition of processing of personal data is very broad, it means “any operation or set of operations such as collection, recording, organization, storage, adaptation or alteration,

retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction".[2]

Broadly speaking personal data may only be processed if:

- The data subject has unambiguously given his consent; or the processing is done in the best interest of the data subject, or
- The processing is done according to a legal obligation; in the public interest, or in the exercise of official authority, or
- The processing is done in the legitimate interests pursued by another party, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject.

From the first condition, the unambiguous consent, one may not infer from the fact, that the person has initially given permission to publish the data that this permission also implies permission to archive these data. It does not. The unambiguousness of the consent implies that it is specific and not by implication.

The third condition, when the processing is done in the legitimate interests pursued by another party, is difficult to use in practice, because these legitimate interests have to be balanced against the interests of the data subject, and there are no fixed criteria to be applied. When it comes to internet harvesting, with thousands or perhaps even millions of persons involved, it is difficult to see how so many people's very different interests may be balanced in one equation.

The only practical possibility is to establish a legal obligation that may authorise the harvesting and subsequent processing of the data, and this was the solution chosen in Denmark.

The data subject has the right to be informed of the processing of his data. However this does not apply where the provision of such information proves impossible or would involve a disproportionate effort.

The data subject also has the right to have incorrect information deleted or corrected. This could have become a serious threat to the integrity of the archive. However, in the explanatory comment to the Legal Deposit Act it was stated that data may not under any circumstances be deleted or changed, but data subjects may have corrections attached to incorrect data.

[2] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Article 2(b). Official Journal L 281 , 23/11/1995 P. 0031 - 0050
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML

LEGAL ASPECTS OF GIVING ACCESS TO THE INTERNET ARCHIVE

The harvesting was not the big issue – except for the fact that the Parliament had to enact a law to authorize it, and that might not be so easy. The big issue was – and still is – how access to the harvested material shall be regulated.

Copyright

The harvesting of internet websites is not confined to the top levels it also comprises the deep web in order to catch publications, reports, articles etc. which may be of interest for a longer period of time, and are important sources for historical research.

The problem of giving the general public online (remote) access to an archive containing this type of material is that such access in many instances will compete with the commercial exploitation of the material. This would be a violation of the three step test, article 9(2) of the Berne Convention, which says that

- It shall be a matter for legislation in the countries of the Union to permit the reproduction of such works in certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author.

In order to avoid a violation of the Berne Convention it was decided that the copyright restriction of access to the internet archive should be as follows:

- The general public may access the internet archive only on site at the legal deposit institutions, e.g. The Royal Library, The State and University Library of Aarhus, and the Danish Film Institute.

- Researchers may have remote online access to material in the internet archive for research purposes, provided the material is not also commercially available.

Data Protection

The internet archive contains all kinds of data, among these also personal data. Most of these data are ordinary, non-sensitive data, like name, address, telephone and information in relation to occupation. Many firms publish information about their employees, e.g. their education and function in the company, photo and contact data. But it has also become common to set up private “family” websites informing about the private life of the family, its history and actual doings, etc. And dating services and other internet meeting points like Facebook may contain very private or sensitive information. These data are normally published by the person or with the person’s consent. But, as we have seen, this consent may not be interpreted to extend to archiving the data. There are also other sources to personal data, like newspapers, electronic journals and books. And both friends and foes may publish “funny” pictures and other pieces of intimate and sensitive information, unthinkingly, not realising that this may be very harmful to the person in question.

As we have seen, personal data may be accessed or “processed” if the processing is done in the legitimate interests pursued by another party, except where such interests are overridden by the interest in the fundamental rights and freedoms of the data subject. This may apply when it comes to ordinary, non-sensitive data. But when it comes to sensitive data, the interests of the general public to access the data are “overridden by the interest in the fundamental rights and freedoms of the data subject”.

It is sometimes presumed that this problem could be overcome by preventing searches for personal names. In many countries, e.g. in the Nordic countries, personal names may be taken from names of locations, a village or some place, in other countries names are taken from professions etc. so in practice this is not possible, unless the system is somehow able to distinguish between personal names and other words. But even if this were possible, you may, as a result of searching for something quite different and non-sensitive, happen to come across sensitive personal information.

So, because sensitive personal data are not separated from other data, it was necessary to declare the whole intrent archive to be “sensitive”, and to deny the general public access to the database. This does not mean, that the internet archive is a “dead archive”, but sensitive data may only be accessed for the purposes for which they were originally created, or for statistical and research purposes. The consequence of this is that as long as it is not possible to isolate the sensitive data, the whole internet archive may only be accessed for statistical and research purposes.

It is most unsatisfactory that an important part of the cultural heritage like the internet archive should be barred from public access. Therefore the in the explanatory comments to the Danish Legal Deposit Act, it is stated that the legal deposit institutions will try to separate sensitive personal data from other data. The assumption is that public websites of government and public institutions and firms and associations of some size are normally handled by professionals and that sensitive information is not made public on such sites. The idea is to separate websites of this origin from other websites and give access to these according to the rules of copyright. It is not perfect, but it may be the best solution so far.

CONCLUSION

Compared to printed publications electronic publications presents us with severe drawbacks in terms of accessibility.

- Since the period of the Enlightenment it has in many countries been possible for a citizen to get access to everything that is printed and published. Much of the material can be taken home on loan, but as a minimum it is available for inspection at the legal deposit libraries. In Denmark this has been the general rule for more that 300 years.

- When it comes to electronic publications, copyright restrictions only permit the general public to access them on site in the legal deposit institutions, while the practical consequences of data protection are, that the general public is denied access altogether, and only researchers may be granted access for research purposes.

It is to be hoped that the last word has not been said on this issue. It seems obvious that the rules concerning protection of privacy need to be adjusted. They were originally made in the context of how to protect sensitive personal information contained in the files of administrators, doctors, and social workers, and they are not apt to solve the problems raised by the internet.

The problems of privacy in relation to the internet are serious for those who have their privacy violated – no doubt about that. But the solution is not to lock up the archive while the violation continues on the living net. That is to lock the gate after the horse has run off.

It should also be mentioned in passing, that not much attention has been offered to the possibly even more serious problems regarding privacy that will arise from digital retro-conversion of printed material, e.g. biographies or unpublished archived material. The present European legislation on data protection may put an effective stop to such endeavours in Europe.

Legal obstacles are often under estimated because one tends to forget that they represent clashes of perfectly legitimate interests between different groups in society. We will have to find a proper balance between society's legitimate need to provide access to our cultural heritage and the equally legitimate need for individuals to have their privacy protected. There are no easy solutions. The only way is to negotiate possible solutions that may accommodate the different groups, and in the end it is a political decision where to draw the line.

****

â€œLegal depositâ€ is a legal obligation that requires publishers to deposit a copy (or copies) of their publications within a specified period of time in a designated national institution. The institution is usually a library, and usually includes the National Library.Â Data protection obligations in national laws for the processing of personal data (data that relates to living identifiable persons such as health and other personal information) should be taken into account, since it cannot be assumed that harvested websites are necessarily compliant.Â Electronic publication: problems of archiving and access to archived information: legal deposit, data protection and related topics. Self-archiving / 'green' open access â€“ the author, or a representative, archives (deposits) the published article or the final peer-reviewed manuscript in an online repository before, at the same time as, or after publication. Some publishers request that open access be granted only after an embargo period has elapsed.Â Refers to the right to access and reuse digital research data under the terms and conditions set out in the Grant Agreement. Research data. Refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images.