LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich. Printed in Germany

Legal Deposit on the Internet: A Case Study

by BIRGIT N. HENRIKSEN

The subject of my paper will be legal deposit on the Internet in Denmark. I will be telling you about the system supporting the legislation, rather than the legislation itself. The system has been developed over the last one and a half years at The Royal Library and is still under development.

I will talk briefly about the modernised legal deposit legislation and then focus on the system.

Our implementation has three main categories:

- a public web site with information and different registration forms;
- a non-public part with programmes for fetching the net publications;
- a non-public archive with very restricted access.

I will illustrate my paper by showing examples from the three parts of the system.

In 1997, the Danish legislation on Legal Deposit was modernised and updated. The previous law had been in force for 70 years and covered only printed works. Working on a definition of what was to be deposited, we ended up with two keywords:

„Work“ and „published“ and with the important point „Regardless of medium“. „Work“ was defined as a limited quantity of information which must be considered a final and independend unit.

„Published“ was defined as: when one or any number of copies of the work have been placed on sale or have been otherwise distributed to the public.

When the law was passed, the matching governmental instruction was produced. It was during this stage, that the concept of „dynamic - static“ appeared. This was created partly as an attempt to define and also limit the number of works to be deposited, partly as a measure to satisfy the software

industry, who with success had protested against the deposit of computer programmes. At present only static documents are covered by the law and therefore archived in our system.

Fig. 1

When the law came into force on 1 January 1998, a new web site was established containing information about the new law, its interpretation and a form for notification of monographs. This web site constitutes the public part of the system.

To support the law, a system was developed for retrieving and viewing the reported net publications. This part of the system is non-public. The site and the supporting system is only in Danish and the non-public part of the system is not available outside the library. The following figures contain screen dumps translated from Danish.

Who deposits?

The person in charge of the technical completion of the digital copy by filling out a registration form at our web site: <http://www.pligtaflevering.dk>.

To ensure the awareness of the legal requirements two mailing campaigns were carried out, one to all public institutions (some 2,500 names) and one to

all known Danish electronic journals (some 500 titles). These mailing campaigns have increased the number of the registrations.

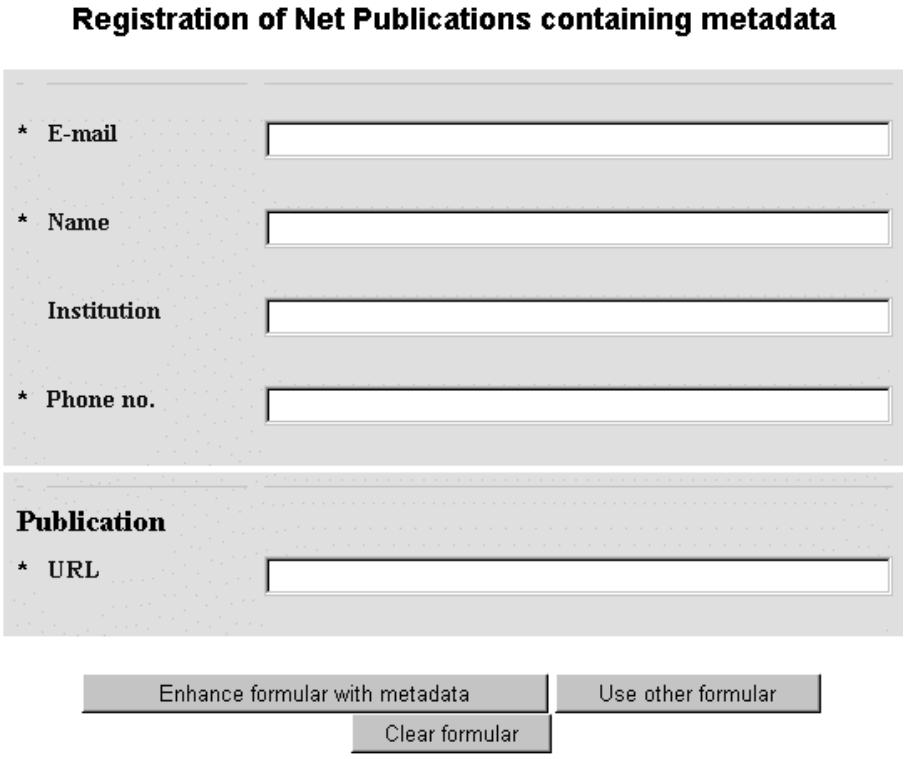


Fig. 2

We have three different registration forms:

- one for monographs with metadata (Fig. 2),
- one for monographs without metadata (Fig. 3) and
- one for periodicals, which require the same input as for ‚monographs without metadata' and additional data about the publication frequency.

The system now supports the use of the Dublin Core Metadata format. Publishers who include the required metadata will have a far easier time when reporting to us than others who have not. They simply add the URL to our site, and we then extract the metadata from the document and the programme

fills in as many of the fields in the registration form as possible. We expect to re-use the extracted metadata for our cataloguing.

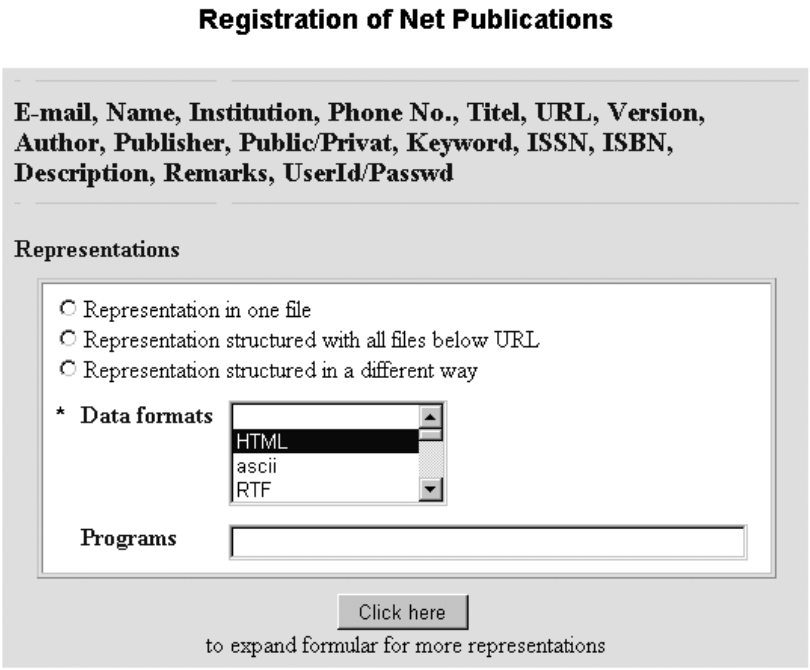


Fig. 3

Fig. 3 shows the registration form for monographs without metadata. Normally, it could take three screens, but I have removed most of the input fields to be able to show the amount of information on a single slide. Information about e-mail, name, institution, phone no, the URL, information about version, author, publisher, ISSN/ISBN, description, keywords, user id. and password for restricted access must be provided to the system. And finally the most important information: For each representation we must know the different file types, the structure of the net publication (one file, files in a tree structure below the URL or something entirely different), information about specific programmes must be available for viewing the publication.

The term ‚representation' has been difficult. A net publication published in e.g. three different formats: HTML, pdf and postscript is considered a single publication but it has three different representations. However we often find through inspection, that only one representation, e.g. HTML is reported to the system.

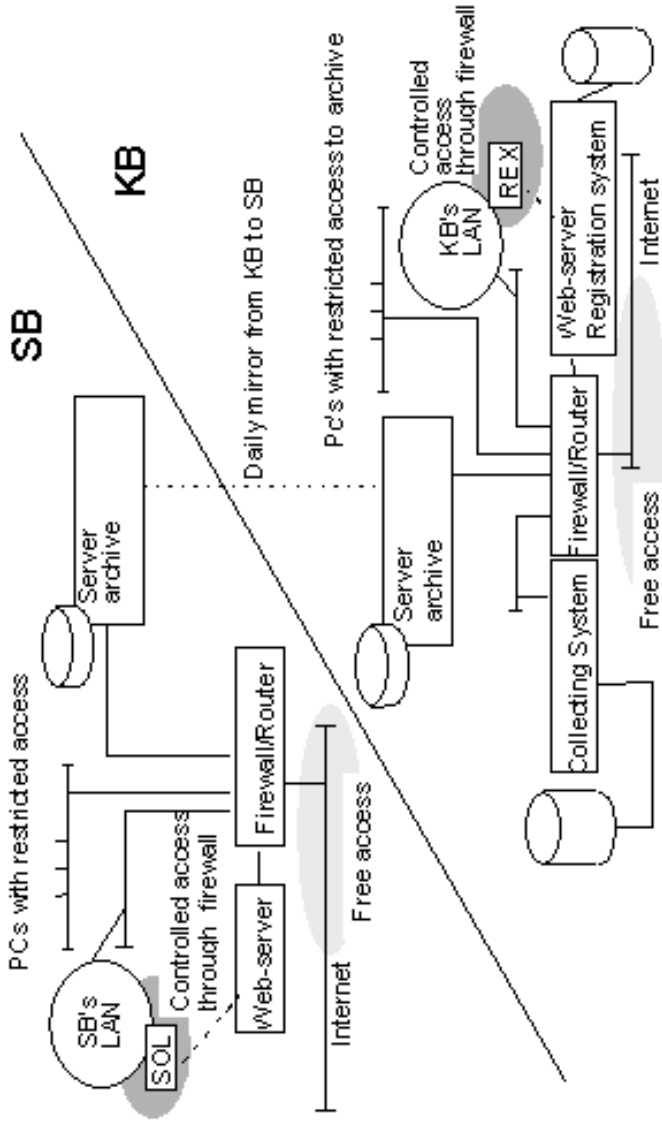


Fig. 4

The Royal Library, Copenhagen (KB) and the State and University Library, Aarhus (SB) are the two institutions involved in legal deposit on the Internet.

Registration system: Software and database containing all reported information about net publications. It is created from information provided in a form on a web server. Here the knowledge about a net publication is born. This is only available at the installation at The Royal Library.

REX/SOL: OPACs of KB/SB with records (in MARC format) on net publications. Records are created in REX and exported to SOL. The OPACs are the roads to the archive. Collecting system: Software and database which handles the fetching and storing of net publications which have been reported to the system. This is only available at the installation at The Royal Library. Archive: Established at KB and mirrored to SB daily. LAN: Internal Local Area Network at KB and SB PCs with restricted access: PCs connected to the LAN in a way that allows it to search the OPAC and reach the archive in order to have the desired document shown, but prevents it from getting electronic copies of documents in the archive. At the moment, only one PC at KB and one PC at SB provide public access to documents in the archive.

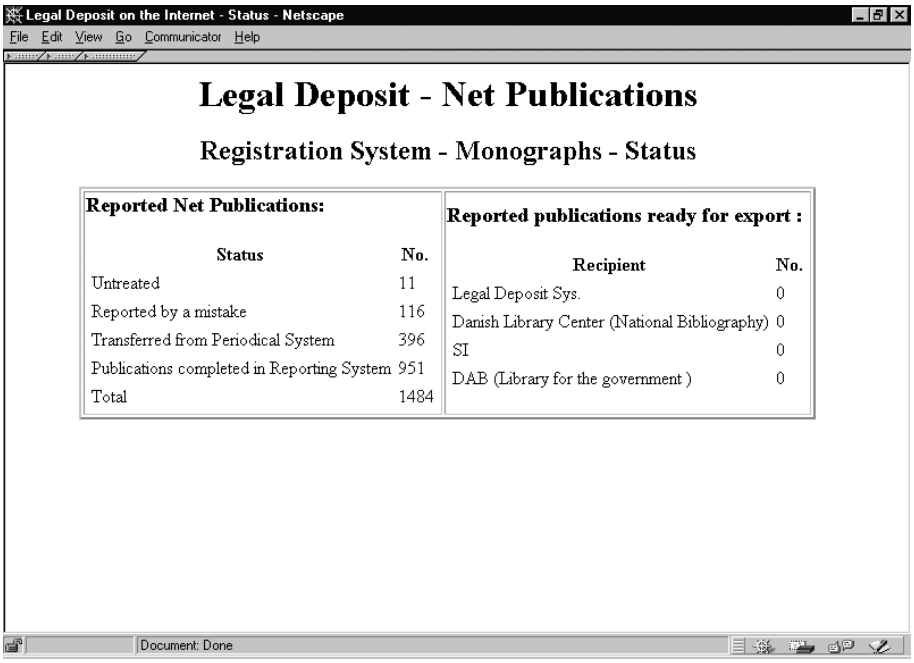


Fig. 5

Fig. 5 shows the status information for the registration system:

The first column shows that as of June 1999 the system contained information about 1,484 registrations. 11 of these have just been reported but are otherwise unprocessed, 116 are reported by mistake and are invalid and will not be fetched and stored by the system, 396 are actually not reported but transferred from that part of the system processing periodicals, and the rest are copied to the collecting system. All figures are links to lists of appropriate registrations.

In connection with the new law the notification structure was simplified due to the fact that the Library informs other relevant institutions of the arrived material. The second column shows the number of registrations for which information has not yet been distributed to the institutions. At the time of this snapshot all information was distributed.

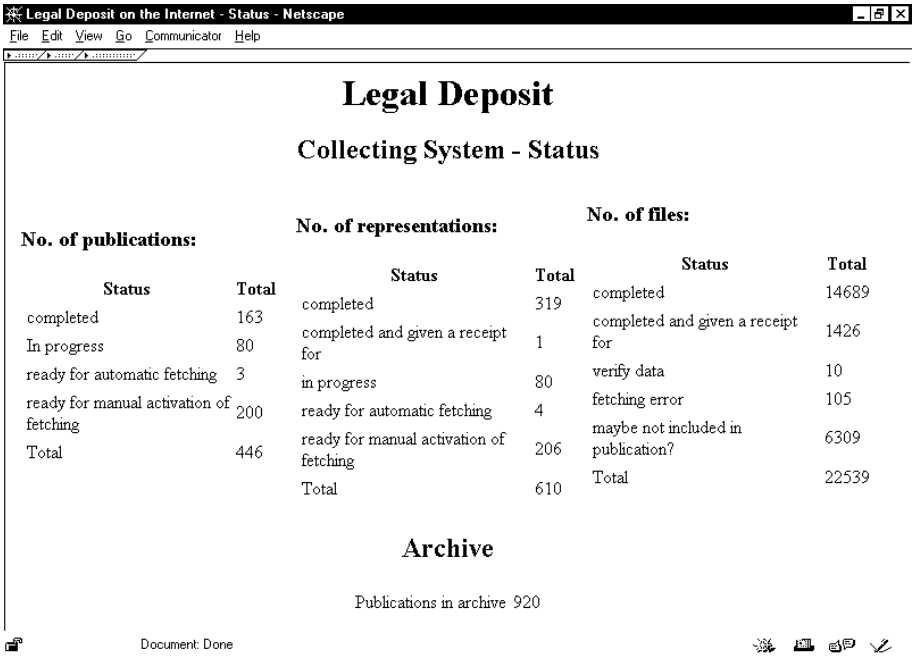


Fig. 6

Fig. 6 shows status information from the Collecting System. The collecting System is a buffer to which the files constituting a publication are fetched and kept until the full publication is fetched and can be moved to the archive.

The status information contains three columns: One for publications, one for representations, where a publication may consist of one or more representations and one for files, where a representation may consist of one or more files. This snapshot shows 200 net publications with a total of 206 representations new to the system and ready to be fetched. They may be manually transferred to the status ‚ready for automatic fetching' and after a short period the fetching will start and the status will change to ‚in progress'. Originally all fetching was implemented as an automatic process, but we realised that very often a full site and not just the publication was reported. This created undesirable work deleting wrongfully fetched files, sometimes up to 25,000 files. We decided to change the system. Now the staff in the Danish Department determines whether a publication is covered by the law and if the answer is ‚yes' they activate the fetching. The fetching is halted if the program is not able to fetch all needed files, or if it cannot determine whether a file is part of the representation or not, or if it cannot verify the file. These situations require manual intervention before the programme continues with fetching. When all the files in a representation have been fetched and verified, as well as all representations in a publication, the full net publication is transferred to

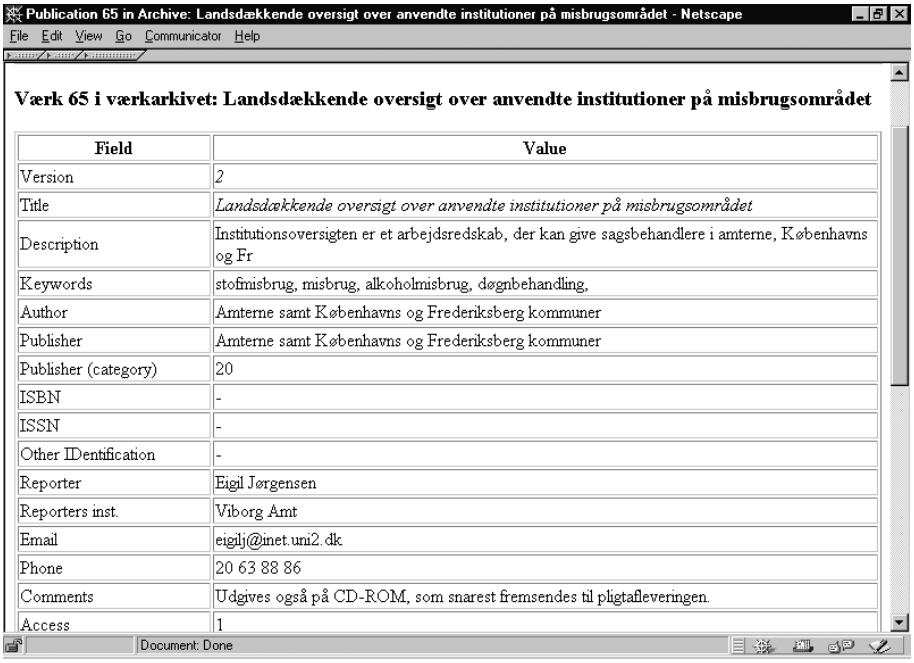
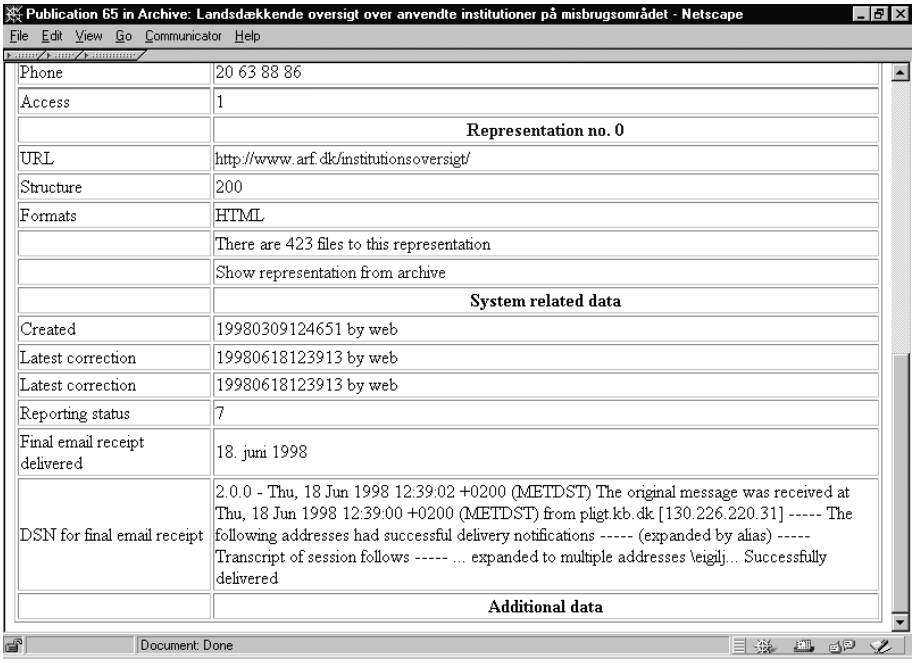


Fig. 7

the archive and mirrored to the State and University Library in Århus with the matching MARC records, which have been made by the staff of the Danish Department at our library. The archive contains at the end of June 1999 952 net publications which consist of 1,299 representations with 104,239 files and a total amount of 1.92 Gbytes. This means that an average net publication consist of 1.4 representations and each representation on average consists of 80 files.

Fig. 7 shows one net publication, publication no. 65, to illustrate the amount of information we store in the database.

First we have version, title, description, keyword, author, publisher, and a code indicating if the publisher is public or private, ISSN/ISBN if available, the reporter's name, institution, e-mail address and phone number, a field for comment and information about access.



Phone	20 63 88 86
Access	1
	Representation no. 0
URL	http://www.arf.dk/institutionsoversigt/
Structure	200
Formats	HTML
	There are 423 files to this representation
	Show representation from archive
	System related data
Created	19980309124651 by web
Latest correction	19980618123913 by web
Latest correction	19980618123913 by web
Reporting status	7
Final email receipt delivered	18. juni 1998
DSN for final email receipt	2.0.0 - Thu, 18 Jun 1998 12:39:02 +0200 (METDST) The original message was received at Thu, 18 Jun 1998 12:39:00 +0200 (METDST) from pligt.kb.dk [130.226.220.31] ----- The following addresses had successful delivery notifications ----- (expanded by alias) ----- Transcript of session follows ----- ... expanded to multiple addresses \eigilj... Successfully delivered
	Additional data

Fig. 8

This is followed by information for each representation:the original URL, the file formats and structure, the number of files in this representation and a link

to the archived version of the net publication. The last section consists of system related data: Timestamps for creation, modification and delivery of email notification to the reporter, telling that we have completed the storing. Finally we store the message from the mail system telling us the status for our mail notification - in this sample the message is ‚The following addresses had successful delivery notifications'. The law gives us up to 3 months from the registration date to fetch the publication. Usually we are quicker than one week but in this example you can see that the fetching period took a little over 3 months. This could be caused by problems with fetching some of the files, and sometimes we have to call the publishers and ask them to correct their publication so we, and other viewers, are able to read the full publication.

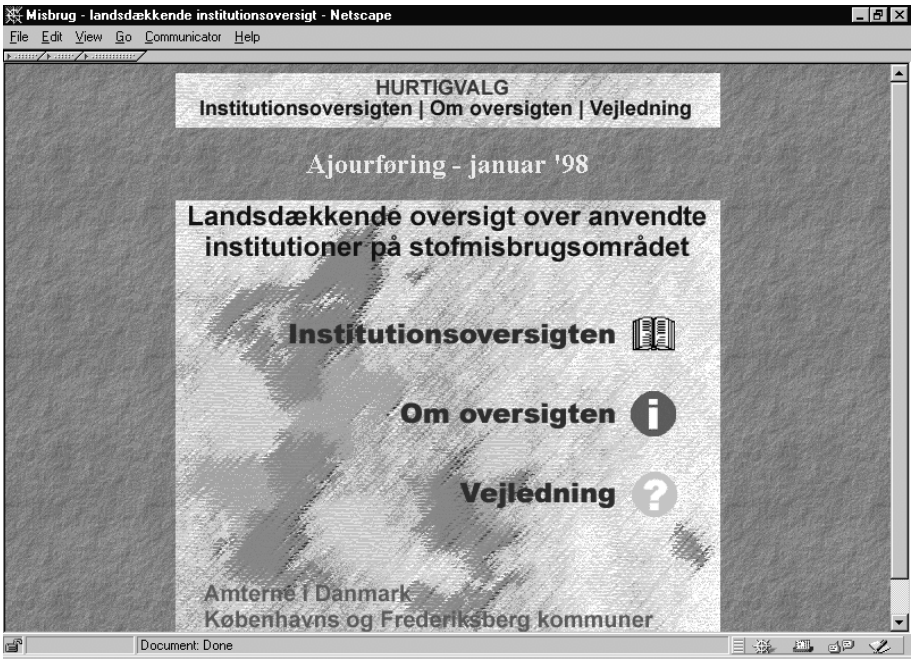


Fig. 9

And Fig. 9 shows the archived version of the net publication. I simply followed the link: ‚Show representation from archive' in the record.

Fig. 10

Fig. 10 shows the result of a search in the OPAC for the word ‚misbrug' or ‚abused' in English.

When the user finds a record in our OPAC (REX) of a net publication, she will then in the near future see two URLs, one pointing to the original publication which can be accessed through the record - provided it is still there - the other will point to the publication on the Royal Library's archive server. At this moment the URL is replaced by a text that informs the user that a copy exists in the archive. If the user is accessing the system from other machines than the dedicated machines in the reading room the user will, when clicking on the URL to the archive, see a text stating that the publication is only available for viewing at a dedicated machine in the reading room at the Royal Library (or at the State and University Library).

These machines, will provide access for viewing and printing, but will not allow any form of digital copying or mailing.

Fig. 11

Fig. 11 shows the original document on the net, pointed to by the URL in the MARC record. You will se that the version on the net is a newer version, released 15 months later, than the one reported to the archive. This illustrates one of the things we have to teach publishers of net publications: that changes to a publication constitute a new version and new versions must be legally deposited too.

This spring we have added a new module to the system which manages the periodicals. I have already mentioned the registration form very briefly and will here show a record for a specific periodical, Hojskolenyt, which is a newsletter. (Fig. 12) The information for monographs is enhanced with information about the publication frequency, which in this example is every Monday in odd weeks. Periodicals are not as easy as monographs to manage and fetch. There are two main problems to be managed in this connection: the fact that publications are not published at really regular intervals, and the fact that the structure of the publishers archives differ. Some publishers choose to overwrite the same URL for every new issue and some create a new URL for

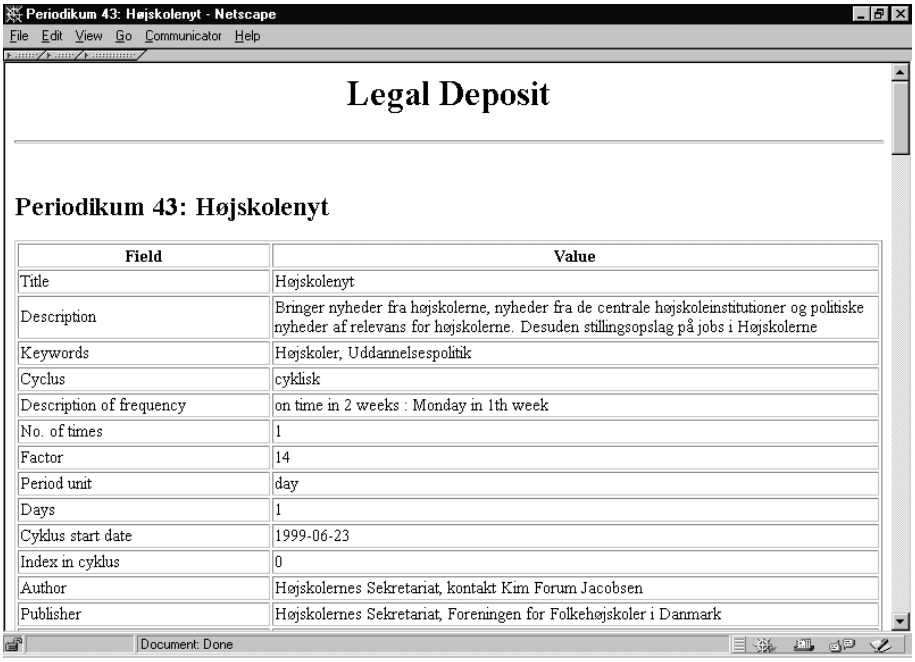


Fig. 12

every issue, thereby creating the need for individual management. In order to minimise the manual work involved in the archival process, the library occasionally makes special agreements with the publishers so that the different issues are placed in an additional structure, just for the library. There is a trend for periodicals which means that they become more like dynamic home pages, thus bringing them outside the scope of the legal deposit legislation.

For each issue there is information about publication ID, timestamp and volume, as well as a link to the archived issue. (Fig. 13)

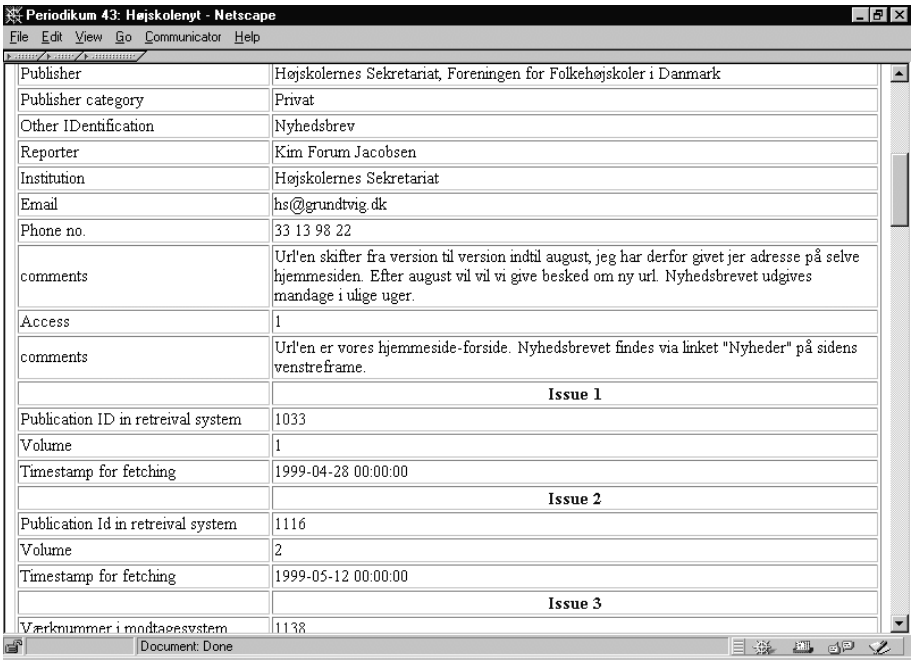


Fig. 13

We have just added a third module to the system: a module handling browser plug-ins. (Fig. 14) This is where information about the name of the plug-in, the version and the platform is maintained and where the plug-in itself is stored. The idea is to keep track of which plug-ins are used in connection with which reported and fetched net publications. Secondly the archive also readily provides the plug-ins, when a new PC in the reading room is installed and configured for viewing the archive.

One of the duties of the Royal Library is to collect, store and make the files available now and in the future. Before the end of the year 2000 the Library will have a plan for long time preservation of this archive. This could be problematic, but if you look at this overview (Fig. 15), you will see that 91 % (84,500) of the files are HTML- or GIF-files and 99 % (91,800) of the files are HTML-, GIF-, JPG- and PDF-files. This means that 99 % of the files are in generally-known and wide-spread formats which we must expect will be maintainable and available in the future.

Field	Value
Name	*PDF*
Version	*4.0*
Description	Plugin til håndtering af PDF dokumenter
	Source no. 0
filename	D:\PLUGINS-SRC\rs32e301.exe
Description	PDF 4.0 for NT
Content-type:	application/octet-stream
Download here:	**D:\PLUGINS-SRC\rs32e301.exe**
	Filetype no. 0
Mimetype	application/pdf
Status	-
	System related plug indata
Commited to the system	19990406164518 of webprogram
Latest correction	19990406164518 of webprogram
	Additional plugin data

Correct plugin | Create plugin | Delete plugin

Fig. 14

In the autumn the system will - on an experimental basis - be enhanced with facilities for managing Danish newspapers on the internet. However, these facilities will bring about a change in the philosophy of how publications can be entered into the system. For internet newspapers we will allow the publisher to provide an electronic copy rather than fetching it ourselves. The change is needed, since the net publisher only stores a soft copy of the publication for a short time, due to resource requirements. This is in sharp contrast to the 3 months ‚period for download' mandated by the legal deposit legislation. In the year 2000 we will analyse the many problems regarding deposits of entire databases.

MIME-type	Collecting system	Archive	Files total
application/msword	64 (0.22%)	169 (0.27%)	233 (0.25%)
application/octet-stream	5 (0.02%)	212 (0.34%)	217 (0.23%)
application/pdf	308 (1.04%)	802 (1.27%)	1110 (1.20%)
application/postscript	1 (0.00%)	30 (0.05%)	31 (0.03%)
application/rtf	1 (0.00%)	5 (0.01%)	6 (0.01%)
application/x-envoy	0 (0.00%)	14 (0.02%)	14 (0.02%)
application/x-pointplus	0 (0.00%)	2 (0.00%)	2 (0.00%)
application/zip	9 (0.03%)	96 (0.15%)	105 (0.11%)
audio/midi	25 (0.08%)	4 (0.01%)	29 (0.03%)
audio/x-wav	26 (0.09%)	12 (0.02%)	38 (0.04%)
image/gif	13301 (44.76%)	22297 (35.42%)	35598 (38.41%)
image/jpeg	2321 (7.81%)	3996 (6.35%)	6317 (6.82%)
image/x-xbitmap	5 (0.02%)	18 (0.03%)	23 (0.02%)
multipart/x-zip	0 (0.00%)	4 (0.01%)	4 (0.00%)
text/css	9 (0.03%)	27 (0.04%)	36 (0.04%)
text/html	13527 (45.51%)	35145 (55.79%)	48653 (52.50%)
text/plain	111 (0.37%)	113 (0.18%)	224 (0.24%)
video/mpeg	0 (0.00%)	1 (0.00%)	1 (0.00%)
video/x-pn-RealVideo	0 (0.00%)	12 (0.02%)	12 (0.01%)

Fig. 15

Birgit N. Henriksen
Royal Library,
PO Box 2149,
1016 Copenhagen, Denmark
bnh@kb.dk