

Setting up a Deposit System
for Electronic Publications

The NEDLIB Guidelines

Johan Steenbakkers

Koninklijke Bibliotheek

How to preserve digital publications

The NEDLIB Guidelines

Setting up a Deposit System for Electronic Publications

The NEDLIB Guidelines

Johan Steenbakkens
Koninklijke Bibliotheek

Production Management Bureau D'Arts, Amsterdam

Design Herlof Schürmann, Amsterdam

Printed by Mart.Spruijt, Amsterdam

NEDLIB Report Series Editor Titia van der Werf

Available from Koninklijke Bibliotheek

P.O. Box 90407

2509 LK The Hague

e-mail info@kb.nl

Title Setting up a Deposit System for Electronic Publications. The NEDLIB Guidelines

ISBN 906259149-3

Author Johan Steenbakkens/Koninklijke Bibliotheek

Date November 2000

Copyright NEDLIB consortium

This study is jointly funded by the European Commission's Telematics for Libraries Programme.

This is report 5 of the NEDLIB Report Series.

content

Preface The proof is in the eating	1
1 Management summary	3
2 Towards a deposit of electronic publications	5
3 Classes of electronic publications	7
4 A paradigm for the deposit system	9
5 Reference Model for an Open Archival Information System (OAIS)	11
6 The NEDLIB Model for the Deposit System	13
7 Storage issues	15
8 Preparing electronic publications for deposit	17
9 Long term preservation	19
10 Setting up a deposit system	21
Annexes	
1 The partners in NEDLIB	23
2 List of NEDLIB reports	25

Preface: The proof is in the eating

Early in 1998 the COBRA-group of national libraries,¹ supported by the Telematics for Library Programme of the European Commission, initiated the NEDLIB project² in order to explore joint solutions for the management and maintenance of electronic publications. NEDLIB stands for 'Networked European Deposit Library'. The partners in the project were eight national libraries, a national archive, two ICT organisations and three scientific publishers. The project ran from January 1998 to December 2000.

The purpose of NEDLIB was to jointly investigate technical and managerial issues concerning the realisation of a depository for electronic publications. A complex issue such as digital archiving could better be tackled by joining efforts and sharing costs. Besides delivering valuable research results, NEDLIB has provided a European forum for consensus building and not in the least, it has promoted national libraries and national archives as key players in the field of digital archiving.

joining
efforts and
sharing
costs

A key role of national libraries is to act as a deposit library. Although the scope and organisation of the deposit library differ amongst national libraries, in essence their aim is similar: collecting, describing, maintaining publications and giving access to them. More generally a deposit library can be described as a collection of publications with the special function of last resort, meaning preserving and guaranteeing access to this cultural heritage of a country, now and in the future.³

In the past few years the depository task of national libraries has, gradually but inevitably, expanded to include electronic publications. A description of this development in the Netherlands is given by Johan Steenbakkens.⁴ A survey, compiled by Libly Martin⁵ in 1999, presents an overview of the management of networked electronic publications by national libraries in various countries. The survey gives a status report for each institution and explains the differences in approach of several national libraries. The 'NEDLIB Local Situations' document⁶ provides a regularly updated overview of regulations, practices, infrastructures and actions at the national libraries and the national archive participating in NEDLIB.

Without exaggeration it can be stated that the realisation of a deposit of electronic publications is currently the major challenge facing deposit libraries and archives. The problem to be addressed is complex because it involves a variety of aspects. These aspects include the policy for the deposit, legislation, the agreements with publishers, the implementation of the organisation and of the technical facilities.⁷

realisation
of technical
facilities

The last aspect mentioned, the realisation of technical facilities, is a moving target as it is closely interwoven with the fast changing Information and Communication Technology (ICT). This concerns especially the long-term access of electronic publications.

1 Computerised Bibliographic Record Actions (Cobra) is a concerted action involving national libraries and bibliographic agencies in Europe.
url: <http://www.kb.nl/gabriel/cobra/>

2 nedlib stands for 'Networked European Deposit Library'. For more details about the project aims, duration, partners and results please refer to:
url: <http://www.kb.nl/nedlib/>

3 Strategic planning in national libraries, by Patricia Donlon and Maurice Line; in: *Alexandria*, 4(2), 1992, pp.83-94

4 Developing the Depository of Netherlands Electronic Publications, by Johan Steenbakkens; in: *Alexandria*, 11(2), 1999, pp. 93-104.

5 Management of networked electronic publications: a status report for various countries, by Libly Martin.
National Library of Canada, December 31, 1999

6 NEDLIB Local Situations – NEDLIB project document – by José Luis Borbinha, Fernando Cardoso, July 2000
URL: http://www.kb.nl/nedlib/results/local_situations_v2.htm

7 For legislative aspects see: Copyright aspects of preservation of electronic publications – Study commissioned by the National Library of the Netherlands, Institute for Information Law, February 1998. URL: <http://www.ivir.nl/Publicaties/koelman/Kbeng2.doc>

For agreements with publishers see: Statement on the development and establishment of codes of practice for the voluntary deposit of electronic publications. Conference of European National Librarians (CENL) / Federation of European Publishers (FEP). Draft, April 2000.

To face this challenge, the best thing to do for libraries is to define the problem in such a way that the ICT-sector can design solutions and subsequently implement them in practice. The experience in the NEDLIB project showed, that if the complexity of the problem as such is emphasised too much, this results in extensive and prolonged discussions. Another effect is that national libraries and archives tend to postpone acting now and wait and see instead of starting to set up their depository for electronic publications and gaining valuable experience.

To avoid this paralysing effect, one just needs to recall the state of the art in preserving printed publications and the conclusion comes to one's mind that the problem of preserving traditional print publications and giving access to them 'for ever' has not been solved. And yet, institutions responsible for this task have built large and valuable deposit collections of printed publications. They have successfully developed facilities, procedures and good practices to maintain these collections as long as possible.

paralysing
effect

pragmatic
approach

Taking this into consideration, the best thing to do is to apply the same pragmatic approach to electronic publications. The strategy NEDLIB recommends libraries and archives to follow reads: 'start working on your deposit of electronic publications right away, obtain facilities and develop the necessary practices'. This is likely a better strategy than only discussing and studying the problem. After all the proof of the pudding is in the eating. 'Obtain facilities' need not necessarily mean to buy ICT equipment, it can also mean to rent the functionality needed. However, if renting is the chosen solution, it is still important for the librarian and archivist to have basic insight in the workflow processes, the functional and technical requirements of a deposit of electronic publications.

The focus of NEDLIB was on technical and functional issues and on good practices for archiving, management and maintenance of electronic publications. Activities were undertaken to develop a generic process model for a deposit system for electronic publications. Available best practices and techniques for archiving and managing specific kinds of electronic publications were mapped to this model. The software solutions and procedures used at the various institutions were analysed and their functionality tested. The test results were taken into account when detailing the process model.

In this NEDLIB report the project results have been consolidated into practical implementation guidelines. In addition to guidelines about setting up a deposit system for electronic publications, NEDLIB also achieved consensus amongst national libraries and national archives in Europe for a standard approach to digital archiving.

practical
implementation
guidelines

Johan Steenbakkens
Project co-ordinator of NEDLIB

1 Management summary

A key role of national libraries and national archives is to preserve and guarantee access to our cultural heritage as recorded respectively in publications and archival records. Although the details of the depository task vary amongst these institutions, in essence their aim and the practices applied are the same.

In the past years the depository task of national libraries and national archives has inevitably expanded towards digital information. Three years ago a group of national libraries, called COBRA, initiated the NEDLIB project in order to explore a joint approach towards the management and maintenance of electronic publications. NEDLIB stands for 'Networked European Deposit Library', a project co-funded by the European Commission.

follow a
common
approach

The expectation of the NEDLIB partners was that, despite differences in deposit library policies and despite different national and local situations, it would be possible for deposit libraries to follow a common approach and to develop a generic high-level design for their deposit system for electronic publications. It was also expected that the results might be to a large extent applicable to national archives as well.

The key objective of NEDLIB was to produce a common model and terminology for the deposit system for electronic publications. On the basis of the Reference Model for an Open Archival Information System, an ISO standard, NEDLIB developed a Process Model for Deposit Libraries. The NEDLIB model describes in detail all processes for handling electronic publications, ranging from ingest, through archival storage and preservation, to access.

A practical step in realising a depository for electronic publications is the implementation of a deposit system: the digital stacks. A basic requirement for this deposit system is that it should support procedures needed to preserve the electronic publications and keep them accessible through time. To implement such a system, specific standards, technologies and procedures are needed. NEDLIB has investigated the availability of these. Some technologies are missing, in particular in the area of digital preservation. NEDLIB has taken first steps to encourage applied research in this area.

digital
stacks

self-
contained
system

An important conclusion of NEDLIB is that a deposit system has to be designed as a separate entity within the digital library environment. There are a number of reasons why the deposit system should be a self-contained system, especially designed for the storage and preservation of electronic publications. The acquisition, cataloguing, search and retrieval functions are provided by other (library) systems and are not as such part of the deposit system. The deposit system of course needs to have appropriate interfaces to integrate the system fully into the digital library environment.

Preservation of electronic publications involves specific methods and techniques. The purpose is to keep the electronic publication 'healthy' and accessible as time goes by. This means preserving the bit-stream and the de-coding mechanism needed to access the information. During the NEDLIB project a first experiment was performed with one of the possible techniques (emulation) for long term preservation.

preserving
the bit-stream
and
the de-coding
mechanism

2 Towards a deposit of electronic publications

A deposit library is an institution where publications are deposited in order to be permanently archived under special conditions. This is done with the aim to guarantee access to the publications, now and in the future. (The same definition can be applied to an archival institution if one reads ‘documents’ instead of ‘publications’. This is also in general the case for the project-results presented in this report.)

deposit of printed publications

In most countries there is a deposit of publications, usually operated by the national library. In fact one should talk about a deposit of publications printed on paper or in short printed publications. In these deposit libraries the prerequisites for operating such a deposit of printed publications have been met. The prerequisites in place are: a policy and selection criteria defining categories of publications to be included in the deposit, legislation or agreements with publishers for depositing the publications, financial and organisational frameworks to support the depository task. Last but not least, good practices have been developed and are applied in order to guarantee preservation and long-term access to the printed publications.

As the depository task extends towards electronic publications all the prerequisites, mentioned above for acquiring and preserving printed publications, also have to be met for this new type of publications. For electronic publications however new prerequisites have to be fulfilled in addition to the conditions mentioned before, namely the ICT-infrastructure and the procedures for storage and preservation of the electronic publications have to be set up. These technical prerequisites have been the focus of the project NEDLIB, leaving aside all the other aspects that of course need to be fulfilled as well, in order to realise a deposit of electronic publications. Examples of good practices for these aspects might be found elsewhere, such as for instance voluntary arrangements, legislation and organisation .

new prerequisites for electronic publications

technical issues of a deposit

So NEDLIB has focused on the more technical issues of a deposit of electronic publications. A generic approach was chosen in order to make the results widely applicable. A generic approach implies that NEDLIB did ignore the great variety of local conditions of the deposit libraries, arising from choices made to meet the prerequisites of a deposit regime. At the same time within the project practical choices were made in order to scope the NEDLIB-project activities. For mainly practical reasons certain classes of electronic publications were taken into account whereas others were not. For the electronic delivery of deposits for example, both web harvesting and offline delivery of CD-ROM publications have been tested.

However it should be stressed that, in practice, the classes of electronic publications chosen for the NEDLIB activities, may or may not be incorporated by a particular national library into its deposit collection, depending on the policy of that library. The same holds true for the different acquisition and deposit procedures. For instance, whether publications are to be acquired from depositing publishers only or whether the deposit collection also includes web harvesting, is to be decided by each national library individually. This is also the case for the relevance of other examples and functions described in this guide. In the end the national library is responsible for defining in detail its own prerequisites for the deposit of electronic publications. In this guide a good starting point is offered for designing the technical conditions for a deposit system for electronic publications, including the procedures connected.

3 Classes of electronic publications

Electronic publications exist in a large variety and for practical reasons they can be classified in classes.

One way of classifying electronic publications is in hand-held or offline (e.g. CD-ROM) and networked or online (e.g. electronic journal).

offline and
online

An offline publication is for obvious reasons also called a hand-held electronic publication. As a rule an offline publication is distributed to the reader in a box and is accompanied by text on the box or in a booklet with instructions how to handle the CD-ROM and with an explanation of the content. (Sometimes the CD-ROM is distributed as an annex inserted in a printed book.)

An online publication is distributed to the reader through a network. (Online publications are occasionally delivered on a hand-held medium to the organisation that will install it for online access.) Distinguishing electronic publications on basis of the distribution medium is useful to analyse the required acquisition procedures and related functions. In the end however this distinction is not relevant. Because no matter on which carrier it has been delivered, in order to preserve the electronic publication for the long-term, the deposit library has to transfer the published information into the controlled storage area of the deposit system.

professional
and
occasional
publishers

A practical distinction is the one between electronic publications of professional (mostly commercial) and of occasional publishers. As a rule agreements with professional publishers about depositing publications will be easier to make. The arrangements might concern the procedures of delivery, general and technical information (metadata) about the publication, such as file formats etc. Publications of professional publishers, depending on the library's deposit policy of course, might comprise a substantial part of the material of the deposit library. To acquire the publications of occasional publishers they have to be either individually searched for or acquired through a publicity campaign alerting the occasional publishers of the possibility (or necessity) to deposit their publication. In this campaign the occasional publishers can also be advised about the procedures and preferred formats to deposit their publications.

Another practical distinction that can be made is between digitally born publications and digitised publications. This distinction is not really of importance within the context of NEDLIB because in essence it has no impact on the technical or functional requirements for a deposit system. Nevertheless it is good to bear these classes in mind for two reasons. If the library digitises publications, it can make its own choices of the image resolution and formats used. These parameters have a great impact on the storage capacity required (costs involved) for keeping the digitised publications. Also for policy reasons it might be useful to distinguish between these classes of publications, for instance for the copyright aspects of giving access.

digitally
born and
digitised

If the content, structure and layout of an electronic publication are defined on beforehand we can classify it as a static publication. If a publication is composed at the moment it is retrieved ('on the fly'), it is called a dynamic publication. One might suggest that a dynamic publication is the result of a search in a specific database. However it makes more sense that not the arbitrary search-result but the database itself together with the basic mechanism for compiling the search-result, is considered to be the (dynamic) publication. In this view both the database and the basic mechanism has to be archived.

static and
dynamic
publications

A relatively new phenomenon is the Internet. The Internet is an international network through which publications can easily be distributed. Besides a tool for distribution the Internet is a powerful publishing environment within which new kinds of publications evolve. Publications on the Internet show a variety in structure and size. By their nature many publications on the web are linked to each other. This implies that in the process of collecting a web-based publication for the deposit, its boundaries somehow have to be defined. To be able to do so, one has to overcome the complication that within a publication the components e.g. paragraphs of an electronic book, are also connected through links. Somehow a mechanism has to be created to distinguish automatically between internal and external links.

web-based
publications

snapshots of the internet

As some countries already have started to do, one might collect occasionally or regularly the information from larger or smaller parts of the Internet (web-archiving) and deposit it as a publication that is worth preserving for future use, for instance all web-sites with a country's extension.

The reader might have noticed that for the classification of electronic publications, the nature of content has not been taken into account. From a perspective that they might require different acquisition procedures and facilities for access, a distinction according to content might be useful. However from a technical point of view the content of the electronic publications is not relevant for defining and realising a deposit system. The underlying assumption is that an appropriately designed deposit system will be fit for every electronic publication whatever the nature of its content.

4 A paradigm for the deposit system

the stacks for digital collections

A system to handle, store and maintain publications in an appropriate way is an indispensable prerequisite for a deposit of electronic publications. To think about a deposit system, its functions and its position within the library or archive organisation, it is helpful to use a paradigm. It might be obvious, but the deposit system for electronic publications should in essence be seen as the stacks for digital collections.

This stack though is not just an ordinary one, but is a highly conditioned and well controlled stack that is especially designed and managed to maintain the deposited publications and to keep them accessible 'forever'. In this respect the deposit system for electronic publications is fully comparable to the actual physical stacks of national libraries and archives where printed books, manuscripts and other documents are kept. Physical conditions, such as temperature, humidity, finish of surfaces, lighting-level and transport mechanisms, are carefully controlled for the sake of preservation. Access to the stacks is well guarded in order not to lose any publication. Before they are stored in the stacks the publications are checked, registered and described in the library's catalogue. The publications are maintained by binding or restoring them and in some cases eventually by transferring the information to another carrier (e.g. microfilm) in the case of deterioration of the original carrier (e.g. paper).

comparable
to the physical
stacks

In the same way a national library or archive needs to have dedicated stacks for storing and maintaining its deposit collection of printed publications, it needs a specific system for storing and maintaining its deposit collection of electronic publications. Procedures for handling electronic publications have to be developed. Special measures have to be taken in order to maintain the integrity and authenticity of electronic publications but also to guarantee accessibility through time.

plan the storage capacity

Just like for the stacks for printed publications, a deposit library has to plan the storage capacity needed to store the amount of electronic publications that will be deposited over a certain period. This is a difficult exercise. In chapter 7 some information is given that is helpful to estimate the storage capacity needed for different types of publications. Of course the amount of storage is directly related, on one hand to the production by the publishers, and on the other hand to the policy defining the content of the deposit collection.

5 The Reference Model for an Open Archival Information System (OAIS)

As a result of the generic modelling approach chosen by NEDLIB a common terminology developed to describe a deposit system of electronic publications.⁸ This facilitated the exchange of ideas and experiences amongst the partners. During this work NEDLIB came across a draft of the Reference Model for an Open Archival Information System (OAIS), an ISO standard under development by CCSDS.⁹ The OAIS model was compared with the modelling work of NEDLIB at that time. The conclusion was drawn that the functions of a deposit system for libraries and archives could be appropriately mapped onto the OAIS model. OAIS was therefore chosen as the basis for the NEDLIB model of the deposit system for libraries and archives.

generic
modelling
approach

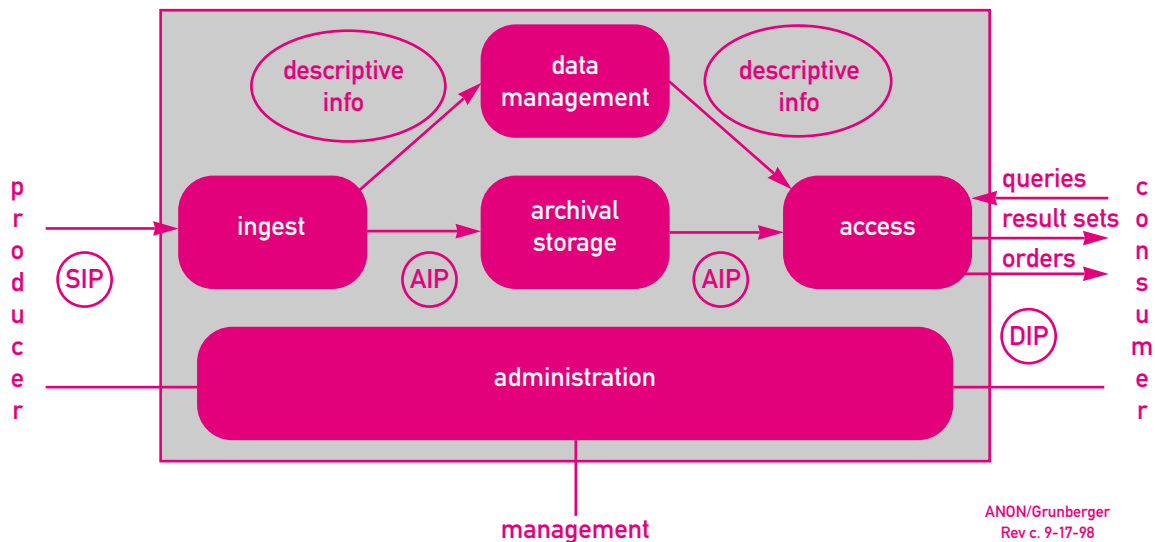


Figure 1 The Reference Model for an Open Archival Information System (OAIS)

In OAIS terminology the handling process of an electronic publication can be described as follows. A publication is prepared for submission to the deposit system and packaged into a SIP (Submission Information Package). Once in the deposit system, it is prepared for storage and repackaged into an AIP (Archival Information Package). When requested for access, it is prepared for delivery and repackaged into a DIP (Dissemination Information Package).

function for
long-term
preservation

In the initial OAIS model, as it is represented in the figure above, five main functions are distinguished: Ingest, Archival Storage, Data Management, Access and Administration. NEDLIB has proposed to extend the OAIS model with a main function for long-term preservation. This proposal has been accepted by CCSDS.

⁸ NEDLIB List of Terms – NEDLIB REPORT 7 – by Genevieve Clavel-Merrin, Den Haag 2000.

URL: <http://www.kb.nl/nedlib/results/NEDLIBterms.html>

⁹ Referencing Model for an Open Archival Information System (OAIS), Don Sawyer / NASA and Lou Reich / CSC.

In the course of project NEDLIB the following subsequent versions of the OAIS Reference Model have appeared and been used as a basis for this NEDLIB report: White Book, Issue 4, September 1998; White Book, Issue 5, April 1999; Red Book, Issue 1, May 1999

URL: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

To comply with the OAIS standard, a library or archive should fulfil a number of responsibilities, listed in the OAIS document. Amongst these responsibilities are: develop a relation with the information providers (e.g. publishers), ensure custody over the information to be preserved, enable the information to be disseminated as authenticated copies of the original or as traceable to the original. **responsibilities**

These responsibilities have been detailed in more detail for scholarly journals by Greenstein and Marcum.¹⁰

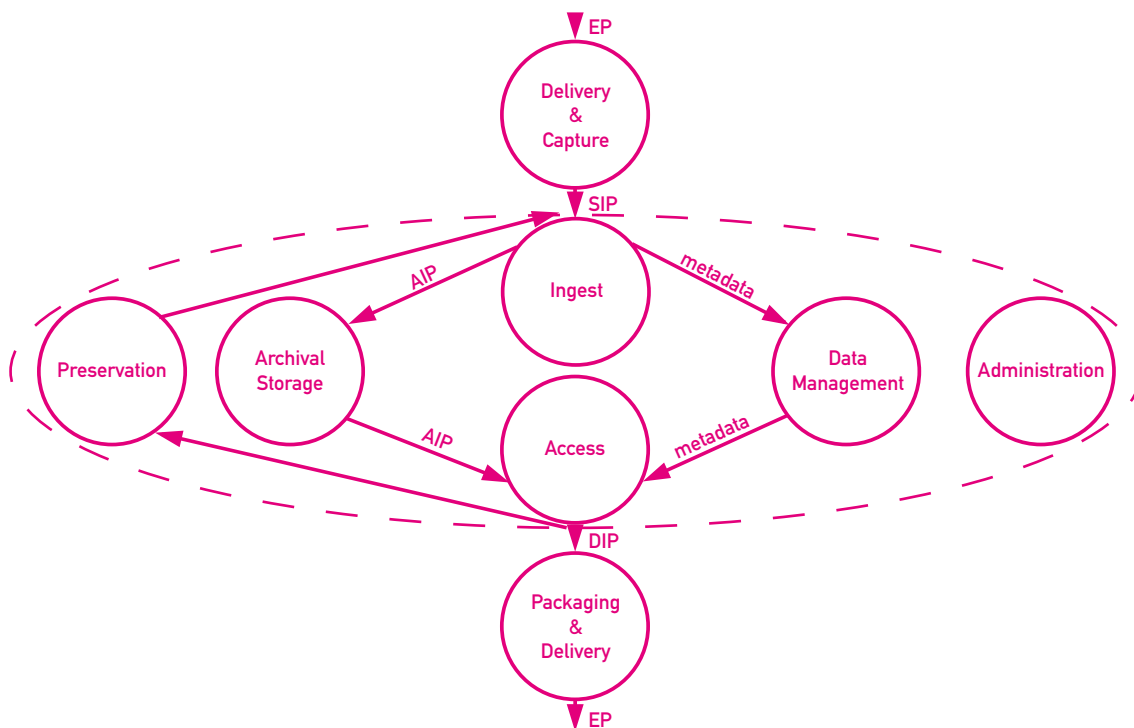
¹⁰ Minimum criteria for an archival repository of digital scholarly journals, by D. Greenstein and D. Marcum, Digital Library Federation, April 2000 – URL: <http://www.clir.org/diglib/preserve/archreq.htm>

6 The NEDLIB model for a Deposit System

a Process Model

On the basis of the OAI model, extended with a main preservation function, the NEDLIB model for a deposit system for electronic publications was developed. The NEDLIB model is a Process Model for the deposit system of a library or archive. The process model is explained in full detail in a separate report.¹¹ In this section a first introduction is given to the model. Figure 2 shows the top-level process of the NEDLIB Process Model for a deposit system.

Figure 2 Top-level processes of the nedlib model for a deposit system based on oais



The top-level processes of a deposit system for electronic publications (DSEP) are described as follows:

1. **Ingest** receives an electronic publication prepared by an interfacing process (Delivery&Capture) and loads it into archival storage.
2. **Archival Storage** takes care of the storage and retrieval of the electronic publications and of the integrity of the bit-streams.
3. **Data Management** takes care of the storage and retrieval of metadata associated with the publication and with systems administration.
4. **Access** makes the archived publication and its associated metadata available through an interfacing process (Packaging&Delivery).
5. **Administration** is responsible for the operation of the system.
6. **Preservation** is responsible for the long-term access and readability of electronic publications.

¹¹ The Deposit System for Electronic Publications: a Process Model – NEDLIB REPORT 6 – by Titia van der Werf, Den Haag 2000.

URL: <http://www.kb.nl/nedlib/results/DSEPPROCESSMODEL.PDF>

The deposit system interacts with the following interfacing processes:

7. **Delivery&Capture** is the input interface of the deposit system.

Delivery&Capture takes care of the pre-processing of electronic publications to be ingested into DSEP. It receives or captures electronic publications and offers a working space for verification, redistribution of data for processing by external systems (Acquisition and Cataloguing) and packaging according to the specifications of a SIP for ingest into the DSEP.

8. **Packaging&Delivery** is the output interface of the deposit system.

Packaging&Delivery takes care of the post-processing of electronic publications retrieved from DSEP. It negotiates access requests, delivers and installs electronic publications, together with appropriate viewing or running software and metadata, for direct access on the library visitor's workstation, via the library access systems or via external systems.

Acronyms used:

DSEP: Deposit System for Electronic Publications; **SIP**: Submission Information Package; **AIP**: Archival Information Package; **DIP**: Dissemination Information Package; **DLS**: Digital Library System.

7 Storage issues

As indicated earlier, a key function of the deposit system is to provide a high quality mass storage facility.

mass storage
facility

disk-
technology
and tape-
technology

Storage technology is a very fast developing area of ICT technology. New technologies are developed and existing ones are improved. A distinction can be made between on the one hand disk-technology and on the other hand tape-technology. Disk-technology provides random access and gives as a rule faster access to the data. Tape-technology provides linear access to the data and is slower. Storage on disks is more expensive than on tapes. The main storage media are Magnetic Disks, Magnetic Optical Disks and Magnetic Tapes. Storage devices can handle a mix of these media. The fast (expensive) media are used to provide quick access from what is sometimes called the 'hot storage'. The slower (cheaper) media are used for the 'cold storage' and are most appropriate for storing the bulk of the deposit data and for backup.

Although storage technology is rapidly developing,¹² storage demand is growing much faster. At the moment the storage devices and connected software will form a substantial part of the costs of a deposit system. It is expected that in the long term this aspect of the costs will diminish relatively, at least if the deposit system is constructed in such a way that replacement of storage technology at appropriate times and by state of the art technology is possible (vertical development of the system). The deposit collection consists of an ever-growing amount of published information - it is by its nature not fixed in size. Therefore it is essential that the capacity of the deposit storage system can be expanded continuously, in keeping with the growth of the deposit collection (horizontal development).

replacement
of storage
technology

expand
continuously

Table 1 Storage load per page

Class of electronic publication	Indicative storage load per page
Internet text page (static)	HTML: 20 Kb (incl. images etc)
Internet page (average storage load)	11 Kb (50% text / 50% images, applications & other)
Electronic journal article	PDF: 50-500 Kb
Digitized publication (text/image)	TIFF: 20-200 Mb
	JPEG: 50Kb – 1 Mb
	JPEG Thumbnail: 10-50 Kb
CD-ROM	Max. 650 Mb

To calculate the storage capacity needed it is relevant to know the storage load of electronic publications. In the case of electronic publications produced by publishers not much can be done to influence the storage load. Of course the storage capacity needed is directly related to the deposit policy, which establishes what kind of electronic publications will be included in the deposit. In the case the library digitises initially printed publications, the library's choice of the format used, can substantially influence the storage load. When calculating the capacity needed, it is essential to bear in mind that as a consequence of the last resort character of a deposit library, at least one backup of all data has to be kept.

backup of
all data

¹² Special section on 'Ultra-High-Density Data Storage' in: Communications of the ACM, November 2000.

URL: <http://www.acm.org/cacm/1100/1100toc.html>

To help estimate the storage capacity, an indication is given in Table 1 of the storage load per page and format. The figures given are derived from NEDLIB experiments and from experience in the practice of NEDLIB partners.¹³ Storage capacity is expressed in bytes using a prefix indicating the magnitude, like Kilobyte (Kb) that stands for 10^3 bytes. For more information, see the table of Storage Capacity Units below.

Table 2 Storage Capacity Units

Term	real Size	Approximation	Example
Kilobyte	2^{10}	$10^3 = 1.000$	10 lines of text (ASCII)
Megabyte	2^{20}	$10^6 = 1.000.000$	50 articles of 10 pages (HTML); 1 - 20 images (JPEG); 20 – 100 thumbnails (JPEG); 2-20 articles of 10 pages (PDF)
Gigabyte	2^{30}	$10^9 = 1.000.000.000$	5 - 50 pages (TIFF) /200 – 2000 pages (PDF)
Terabyte	2^{40}	$10^{12} = 1.000.000.000.000$	200-2.000 journals with 340 articles of 10 pages over a period of 10 years (PDF); 50-500 books of 100 pages (TIFF); 2.000-20.000 books of 100 pages (PDF)
Petabyte	2^{50}	$10^{15} = 1.000.000.000.000.000$	This amount of storage is already occasionally being implemented in industry
Exabyte	2^{60}	10^{18} etc.	Future capacity units have already been named
Zettabyte	2^{70}	10^{21}	
Yottabyte	2^{80}	10^{24}	

¹³ op.cit. 6

8 Preparing electronic publications for deposit

To analyse the pre-processing steps needed for preparing electronic publications for ingest into a deposit system, the analogy with the processing of printed publication made earlier in section 4 might be useful. In fact we shall take the example of an offline or hand-held publication to illustrate similarities and differences.

For the preparation of an offline publication, for instance a CD-ROM publication, facilities for installation are needed. A workstation with the appropriate drivers and readers is required where the new publication can be opened, checked and registered. The reason for checking the publication is to guarantee the quality of the deposit collection from the start. Registration of the publication will be done with the (external) acquisition functionality of the library system, a functionality that is already operational for printed publications. As part of the existing acquisition functionality a claim might be issued if necessary to the publisher, e.g. for missing issues. Similarly the CD-ROM publication will subsequently be catalogued and provided with keywords or subject-headings using existing functionality of the library system. The bibliographic descriptions, the keywords and subject-headings are incorporated into the online catalogue so that the publication can be retrieved in response to a search. In the catalogue a shelf-number is given or a link if the CD-ROM is stored in a jukebox. All of these steps do not differ from established processes for handling printed publications. They need to be performed for all new incoming deposit collection titles, whether they concern printed or electronic publications.

An additional step needs to be taken to incorporate the CD-ROM publication into the deposit system. This non-traditional step involves the transfer of the content from the CD-ROM carrier to the deposit store. In the NEDLIB model this step is supported by the Delivery&Capture functionality. A CD-ROM publication stored in this way can be accessed online through a network. Once a CD-ROM publication has been transferred from its original carrier to the deposit transfer area, it can be treated in the same way as all new incoming electronic publications.

transfer the
content to
the deposit
store

In order to prepare electronic publications for ingest into the deposit system all new incoming publications should first be copied to the deposit transfer area. This is a temporary storage space for clearing incoming deposit publications. It is the pre-processing work area of Delivery&Capture. The electronic publications can then be checked on integrity and completeness. The table of contents file, provided by the publisher in an agreed format, is a good means to verify completeness. Title information provided by the publisher is checked against the catalogue entries. If it concerns a known title it will only be necessary to update the catalogue record, else a new record needs to be created.

checked on
integrity and
completeness

New descriptive metadata accompanying the publication may be copied to the cataloguing system. Automatic pre-cataloguing may be implemented in such cases. More refined content indexing information provided by the publisher may complement the online catalogue entries. For the purposes of access a tag is added to link the deposit copy of the electronic publication to its catalogue record. The National Bibliographic Number (NBN)¹⁴ can be used for this purpose.

link
the deposit copy
to its catalogue
record

In practice it appears that publishers do not always provide all the necessary information and if they do, they do not do so in any standardised fashion. Different publishers may for example provide descriptive metadata in SGML format, but they will do so according to their own, specific, home grown document type definition (DTD) flavour. The pre-processing activities of Delivery&Capture therefore have to be continuously adapted to accommodate publications coming from different publishers. To control this ever-growing variety and optimise metadata exchanges between libraries and publishers it is necessary to agree with the publishers on further standardisation of the delivery formats.¹⁵

¹⁴ Identification, BIBLINK project document, May 1997. URL: <http://hosted.ukoln.ac.uk/biblink/wp2/d2.1/>

¹⁵ Standards for Electronic Publishing: an overview – NEDLIB REPORT 3 – by Mark Bide, Den Haag, 2000. URL: <http://www.kb.nl/nedlib/results/e-publishingstandards.pdf>

To incorporate the electronic publication in the deposit system the publication has to be loaded into the depository store. For this purpose functionality of the Capture&Delivery module is necessary to package the publication data into a SIP for ingest into the deposit system. So eventually every electronic publication will be stored as one data object in the deposit system. At this point the issue of authenticity has to be introduced. Ideally, publications are stored and maintained in the deposit library in the form in which they have originally been published or in their authentic form. But in practice this is not feasible. We have seen already, for example, that we need to transfer offline publications from their original carrier to the deposit store. To be useful the term ‘authentic’ should be further defined. Authenticity criteria for electronic documents should be developed. As Jeff Rothenberg¹⁶ points out these criteria might be different for different classes of publications or different kinds of use of such documents. For example, casual readers will have different demands than scholars or lawyers. For text documents authenticity is partially specified through the formats used (TIFF, PDF or HTML). For multimedia publications this is more complex.¹⁷ Again, by analogy, we should remember that also for printed publications the issue of authenticity is of importance. But in this case a practical approach to the problem has been chosen. For example, newspapers that have been transferred from paper onto microfilm are never doubted on their ‘authenticity’. It might be suggested to follow the same pragmatic approach for electronic publications in order not to be trapped in a theoretical discussion on authenticity.

the issue of authenticity

a practical approach

In table 3 an overview of the pre-processing steps for an electronic publication is given. A detailed description of these steps is given in NEDLIB report 6.¹⁸

Table 3 Pre-processing steps of an electronic publication for loading into the deposit storage system

Process	Steps
Delivery&Capture	Copy EP to the deposit transfer area Check files on integrity and completeness Extract metadata Assign/Get unique identifier of logical publication unit (NBN) Register EP with external systems
Acquisition	Register title of EP
Cataloguing	Create new title record for EP / Update existing title record
Delivery&Capture	Create SIP
Ingest	Receive SIP in staging area Unpack and validate SIP Extract metadata for Data Management Extract EP and repackage into an AIP
Archival Storage	Store AIP
Data Management	Store metadata
Ingest	Clean-up staging area
Delivery&Capture	Clean-up deposit transfer area

¹⁶ Using Emulation to Preserve Digital Documents, by Jeff Rothenberg. Koninklijke Bibliotheek, The Hague, 2000

¹⁷ Best Practices for Digital Archiving: an information life cycle approach, by Gail Hodge, in: D-Lib Magazine, 6(1), January 2000.
URL: <http://www.dlib.org/dlib/january00/orhodge.html>

¹⁸ op.cit. 10

9 Long term preservation

Books, or to be more exact, books printed on paper are pretty durable. Left alone under not too extreme conditions they can still exist after many centuries. Electronic publications however will not survive by accident. A pro-active approach is required, both to keep the mere data available and to preserve the decoding of the data, so as to be able to actually read the publication. In the past years several experts and institutions¹⁹ have tried to raise awareness and warned that preservation of digital information is a problem that can no longer be ignored. A range of activities has been conducted to study the issue and to define the problem. Also a number of initiatives²⁰ have been taken to achieve that electronic publications will remain available, now and in the future.

a pro-active approach

NEDLIB has determined Preservation as a contained function within a deposit system. Contributing to the ISO review process, NEDLIB has proposed to add the function 'Preservation' to the Reference Model for an OAIS. This proposal has been accepted. Preservation is a vital but at the same time very complex function of a digital archive.

Preservation will require specific procedures and techniques. This is obvious if one realises that all electronic publications are in essence encoded bit-streams. To keep electronic publications in good order and accessible as time goes by, it is necessary to preserve the bit-stream together with the decoding mechanism needed to interpret the bit-stream and to translate it into human-understandable information.

As the electronic media carrying the bit-stream rarely survives more than a few years, the bit-stream must be copied before the medium starts to deteriorate or becomes obsolete because of changes in the technology. This problem of medium migration can be solved using today's ICT technology. Even so, large data repositories with mass storage experience, show that ever-increasing data stores require ever growing time frames for medium migration so that consecutive migration cycles are beginning to overlap each other – in other words a new cycle is started even before the previous one is finished. Continuous improvement of medium migration and storage techniques is needed to avoid such problems.²¹

medium migration

Keeping the decoding mechanisms available through time is a far more difficult problem. But saving the bit-stream of an electronic publication without such a mechanism is, in Jeff Rothenberg's words, like saving the hieroglyphics without saving a Rosetta stone – a key to the information saved. The key to the digital information has somehow to be preserved as part of the deposit collection. This is a major challenge that still has to be addressed. For this purpose new standards and techniques have to be developed for keeping the de-coding mechanism available through time and across major changes in ICT technology. Some suggestions have been done that might help to develop the required preservation technology. Jeff Rothenberg has proposed to use the technique of emulation and Raymond Lorie has suggested an implementation based on the use of the universal virtual computer (UVC) technology.²²

preservation technology

Innovative research aimed at the further development of standards and technology is needed before the preservation of electronic publications will be a reality. The ICT market has just begun to respond to this demand for technology that maintains digital information accessible across time, independently of major ICT platform or paradigm changes.

19 The Commission on Preservation and Access produced the documentary film 'Into the future: on the preservation of knowledge in the electronic age.' American Film Foundation, Santa Monica CA 90406. See also: URL: <http://www.clir.org/programs/otheractiv/intro.html>

20 A good overview of initiatives and projects in digital preservation is given by the Preserving Access to Digital Information (PADI) subject-gateway. – URL: <http://www.nla.gov.au/padi/>

21 Avoiding a Data Crunch, by Jon William Toigo, in: Scientific American, May 2000.
URL: <http://www.sciam.com/2000/0500issue/0500toig.html>

22 Long Term Archiving of Digital Information, by Raymond A. Lorie, Research Report RJ 10185 (95059), March 28, 2000

deposits for content and for rendering software

Besides the standards and technology, it will be necessary to realise on the short term deposits both for the content of electronic publications (the bitstream) and for the rendering software (the decoding mechanism).²³

²³ To this end Software Repositories and Representation Information Repositories need to be established. For a first introduction to these concepts see NEDLIB REPORT 6.

10 Setting up a deposit system

Setting up a deposit system requires a major effort and a substantial amount of resources. At the same time the conclusion of NEDLIB is that, looking at the state of the art of technology,²⁴ it can be done. And what is more: waiting longer will not solve the problem for deposit libraries and archives. On the contrary, while standards for producing documents and for publishing are still under development, they can be modulated in such a way that they will accommodate the conditions for long term preservation. This also holds true for the development of storage facilities and the functionality and technology involved. Libraries and archives can contribute to these ICT developments because, as users, they have unique expertise in handling and maintaining information through time.

When setting up a deposit system, it is essential to use the known approach for success in a large enterprise: keep matters simple. The starting point to do so is to consider that, just as the book-stacks are not the library, neither is the deposit system the total digital library. It need not provide total functionality for handling electronic publications. This means that the functionality of the deposit system has to be scoped carefully.

keep
matters
simple



Figure 3 The deposit system for electronic publications

Figure 3 illustrates how the deposit system should on the one hand be a self-contained system, but on the other hand it should be integrated within the digital library environment.

The deposit-system can best be set up as a separate entity within the ICT infrastructure of the library or archive. This might seem obvious, but in the process of developing a deposit system, there is a danger that one will try to incorporate all the functions needed for processing electronic documents and the delivery of this information. This should not be done for several reasons.

avoid a system
that is
too complex

One reason is to avoid a system that is too complex, as such a system will be hard to design. Even if this could be done, it would probably be impossible to find a provider able to deliver the system at terms affordable for libraries and archives.

The second reason is that, as a rule, systems for book-processing and information delivery are already in place within libraries and archives. Both the technological and the functional development of these functions have their own dynamics. The function of information delivery has to keep up with the requirements of the end-users and with the relevant access technology developments. The same holds for the systems for cataloguing, acquisition, circulation etc.

systems
are already
in place

²⁴ Standards for a dSEP: standards for the Implementation of a Deposit System for Electronic Publications (DSEP) – NEDLIB REPORT 4 – by Bendert Feenstra, Den Haag 2000. – URL: <http://www.kb.nl/nedlib/results/dsepstandards.rtf>

The deposit system also has its own pace of development. The deposit system by its nature will have to be permanently upgraded not only in capacity (horizontal development), but also to keep up with the developments of the technology for storage and data handling (vertical development).

the deposit system
has its own pace of
development

integrated
within the ICT
infrastructure

In addition to designing and implementing the deposit system as a separate system, it is essential that it will be appropriately integrated within the ICT infrastructure of the library or archive. As a result the interfaces for input and for output of the system have to be carefully chosen and designed.

The functionality required for the deposit system has in general already been discussed in the previous sections. More details can be found in the NEDLIB Reports and elsewhere in the literature. Of several functions a pretty complete description can be made and used as a basis for ordering (parts of) a deposit system. Much effort is still required before the functionality required for long term preservation can fully be understood and adequately described. Some specific functional requirements of preservation however are already clear. The functionality for refreshing has to be carried out pre-emptively because the deposit is the last resort facility. For the same reason an appropriate backup facility should be an integral part of the design of the deposit system.

For reasons of efficiency the system should by and large carry out all its functions automatically. This is the only way that national libraries will be able to maintain a large, varied and continuously growing collection of electronic publications.

carry out all
functions
automatically

Annex 1 Project partners in NEDLIB

Organisations and persons who have contributed to NEDLIB:

Koninklijke Bibliotheek (The Netherlands)	Johan Steenbakkers Lex Sijtsma Titia van der Werf
Agentschap Rijksarchiefdienst (The Netherlands)	Charles Noordam Adri Vliet Eric Burger Tamara van Zwol
Bibliothèque Nationale de France (France)	Sonia Zillhardt Elisabeth Freyre Catherine Lupovici Julien Masanès Pierre Echegaray
National Library of Norway – Mo I Rana (Norway)	Torstein Olsen Fred Moerman Knut Tore Breivik Hilde Høgås
Helsinki University Library (Finland)	Juha Hakala Inkeri Salonharju Petri T Heliniemi
Center for Scientific Computing (CSC) (Finland)	Kirsti Lounamaa Mika Rissanen
Die Deutsche Bibliothek (Germany)	Hans Liegmann Jörg Berkemeyer
Biblioteca Nacional do Portugal (Portugal)	Fernando Campos Fernando Cardoso José Luis Borbinha Pedro Marques
National Library of Switzerland (Switzerland)	Genevieve Clavel-Merrin Michel Moret Barbara Signori
Biblioteca Nazionale Centrale di Firenze (Italy)	Claudio di Benedetto Giovanni Bergamin
Instituto De Engenharia de Sistemas e Computadores (Portugal)	Nuno Freire
csc Ploenzke AG (Germany)	Jürgen Kessler Markus Fischer
Ecsoft (United Kingdom)	Vony Gwillim Sally Owens Clive Bennett Sue Purves
Sponsors	Kluwer Academic (The Netherlands) Elsevier Science BV (The Netherlands) Springer-Verlag (Germany)

Annex 2 List of NEDLIB project documents

Long-term Preservation

NEDLIB Report 1: An Experiment in Using Emulation to preserve Digital Publications, by Jeff Rothenberg (Den Haag, 2000).

NEDLIB Report 2: Metadata for long-term preservation, by Catherine Lupovici and Julien Masanès (Den Haag, 2000)

Standards

NEDLIB Report 3: Electronic Publishing Standards – an overview, by Mark Bide (Den Haag, 2000).

NEDLIB Report 4: Standards for implementing a Deposit System for Electronic Publications, by Bendert Feenstra, IBM Nederland (Den Haag, 2000).

Guidelines

NEDLIB Report 5: Setting up a Deposit System for Electronic Publications. NEDLIB Guidelines, by Johan Steenbakkens (Den Haag, 2000)

Modelling

NEDLIB Report 6: The Deposit System for Electronic Publications – A process Model, by Titia van der Werf (Den Haag, 2000).

Definitions and concepts

NEDLIB Report 7: List of NEDLIB Terms, by Genevieve Clavel-Merrin (Den Haag, 2000)
Inventory of terms and standards relevant for NEDLIB,
by José Luis Borbinha, Fernando Cardoso and Nuno Freire (Lisboa, 2000)

Deposit Library Requirements

Inventory of local Situations,
by José Luis Borbinha, Fernando Cardoso and Nuno Freire (Lisboa, 2000)

Preliminary Functional Requirements Specifications,
by José Luis Borbinha and Fernando Cardoso (Lisboa, 1998)

First version of a High-level design for a Deposit System of Electronic Publications,
by José Luis Borbinha (Lisboa, 1998)

Tools

Description of Tools for NEDLIB, by Juha Hakala and Joerg Berkemeyer (Helsinki, 2000)

Scorecard for the Functional Testing of NEDLIB Tools, by Lex Sijtsma (Den Haag, 2000)

Nedlib Annual Reports

Nedlib Annual Report 1998

Nedlib Annual Report 1999

Nedlib Annual Report 2000