

STORIES

D-Lib Magazine September 1999

Volume 5 Number 9

ISSN 1082-9873

Long-term Preservation of Electronic Publications

The NEDLIB project

Titia van der Werf-Davelaar
Koninklijke Bibliotheek, National Library of the Netherlands
titia@python.konbib.nl

The NEDLIB project

NEDLIB was initiated by [CoBRA+](#), a permanent Standing Committee of the Conference of European National Libraries (CENL). The project was launched on January 1, 1998, with funding from the European Commission's Telematics Application Programme, and will run till the end of 2000. Eight national libraries in Europe, one national archive, two ICT organisations and three major publishers are participating in the project. The Koninklijke Bibliotheek, National Library of the Netherlands, leads the project.

NEDLIB, which stands for Networked European Deposit Library, aims to develop a common architectural framework and basic tools for building deposit systems for electronic publications (DSEP). The project addresses major technical issues confronting national deposit libraries that are in the process of extending their deposit, whether by legal or voluntary means, to digital works [[ref. 1](#)].

One important piece of work being carried out by the project is the functional specification and overall design of a DSEP. The main objective is to identify functional requirements that are common to all deposit libraries in order to arrive at a "generic" high-level design of a DSEP that can serve as a basis for local implementations by individual deposit libraries.

A common workflow for handling deposited electronic publications was defined and helped to identify common functional requirements. A major step forward in the conceptual design of a DSEP was made in December 1998, when the project consortium agreed to adopt the Open Archival Information System (OAIS) model as a

Reference Model [[ref. 2](#)]. The fact that the model was being used by other, similar, projects such as [CEDARS](#) in the UK and [PANDORA](#) in Australia, prompted the decision. Now work is being carried out to detail a DSEP process model and data model, based on the OAIS framework, applicable to all deposit libraries, and detailed enough to enable consistent implementation design and development work.

The second main objective of the project is to address the issue of long-term digital preservation. Work in this area should provide better insight into the pro's and con's of different long-term preservation strategies as applied to digital deposit collections. The characteristics of electronic publications and other categories of digital deposit material and their associated preservation and authenticity requirements need to be defined. The NEDLIB partners recognise that many aspects, including cost-effectiveness, legal restrictions, agreements with publishers, and user access requirements, ultimately need to be taken into account when policy choices for preservation strategies are set. For NEDLIB, however, the focus lies on the technical issues of preservation. The Koninklijke Bibliotheek has taken a first tentative step to help define and test the technicalities of preservation mechanisms by starting an emulation experiment with Jeff Rothenberg. The first stages of this experiment will be implemented in NEDLIB.

Besides work on abstract modelling and experimental preservation strategies, NEDLIB is very much geared towards producing pragmatic, ready-to-use results. Recommended standards and conventions for technical solutions are documented in order to provide deposit libraries with practical guidelines when implementing a DSEP. Practical experiences, technical infrastructures and organisational approaches taken by individual NEDLIB partners are gathered and compiled in such a way that these experiences can be of use to other libraries.

The third and last main objective of the project is to build a demonstrator system, with tools and software already in use by project partners or developed by NEDLIB, covering all functional aspects of a DSEP. Software and tools are being developed, tested and integrated in functional building blocks of the demonstrator. Existing library systems, such as the online public access catalogue (OPAC) and the library acquisition and cataloguing systems, which are external to, but need to interact with a DSEP, will interface to the demonstrator. During the demonstration stage, the handling of electronic publications from acquisition to access will be demonstrated, with sample material provided by Elsevier Science, Kluwer Academic Publishers and Springer-Verlag.

In this article I will expand a little on the first two work areas: the modelling of a DSEP on the basis of OAIS and experimenting with emulation for preservation.

Modelling of a DSEP on the basis of OAIS

The OAIS document, drafted by the Consultative Committee for Space Data Systems (CCSDS) of NASA, is a technical recommendation prepared for formal review as a draft ISO standard. It establishes a common framework for functional and information modelling concepts applicable to any archive. It is specifically applicable to

organisations that have a responsibility to provide long-term access to digital information.

As such, the OAIS model is relevant to deposit libraries. The prospect that such a model can provide a solid basis for standardisation within digital archives and promote greater vendor awareness and support of archival requirements, was decisive for the NEDLIB partners. They decided to map deposit library requirements to OAIS and to detail the OAIS model into a DSEP model for deposit libraries. In this way, NEDLIB hopes to contribute to the OAIS standardisation work and, ultimately, to promote DSEP implementations conforming to the OAIS standard.

Scope of a DSEP in a digital library environment

In the same way as the OAIS model presents a high-level view of the interaction between the OAIS and the environment surrounding it, it is necessary to position a DSEP in the digital library environment. Which functionality is within scope of a DSEP, and which belongs to the Digital Library System (DLS) as a whole? Most of the functionality relating to the selection and description of digital works, the creation of finding aids (such as bibliographies, catalogues, subject-guides and indexes), and the provision of user access, is part of the broader digital library configuration. Consequently, OAIS functional entities, such as Data Management and Access, which overlap general functional requirements of a digital library, need to be delimited in some detail, for DSEP purposes. Additionally, it is important to specify how a DSEP interfaces with the digital library system. This work is ongoing as part of the consensus-building process in NEDLIB.

Process model for a DSEP

The workflow for handling electronic publications from selection for inclusion in the deposit collection to end-user access has been detailed into a prototype process of 13 steps. This process has been mapped to the OAIS set of functional entities. Figure 1 shows the result of this exercise.

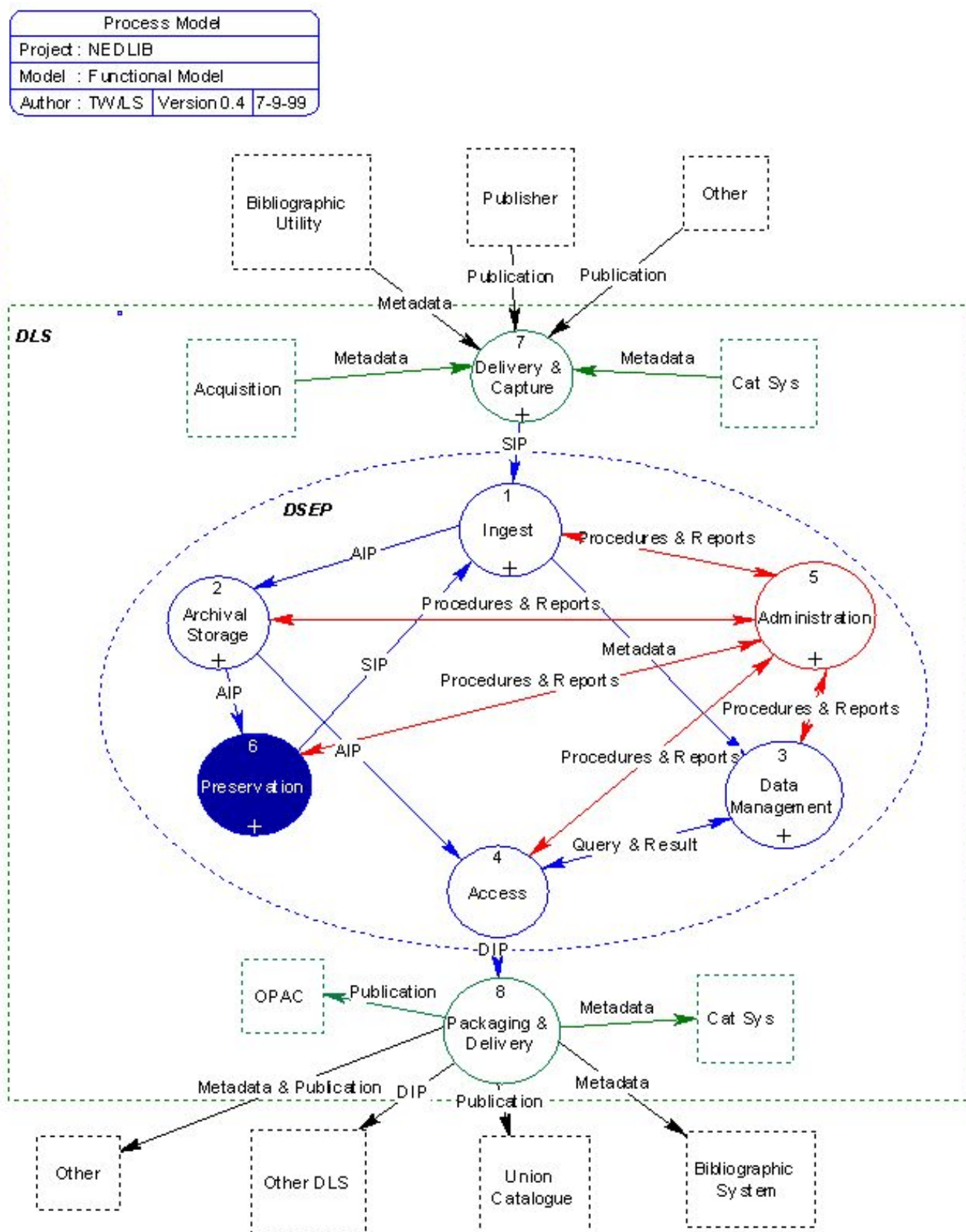


Figure 1. DSEP process model for handling electronic publications

The interfacing modules to a DSEP

A DSEP interacts through 2 interfaces to existing library systems:

(7) Delivery and Capture

The library acquisition system and the associated procedures are responsible for selecting and acquiring the deposit copy of a publication. The procedures may vary with each library, each publisher and each publication type (CD-ROM, Web pages, etc.).

To be able to ingest publications into a DSEP, an interface is needed to ensure the publication is (re-)packaged according to the specifications of a SIP (Submission Information Package). This interface may need to generate, if necessary, accompanying instructional data, in order for the Ingest module of the DSEP to be able to process the publication properly.

This "pre-processing" interface is needed because deposit libraries cannot dictate submission formats to publishers: in principle, they have to accept all formats published on the market.

Most development work at deposit libraries presently concentrates on this interface. It requires much tailoring, as some publishers provide table of contents and others don't, some provide full-text versions for indexing and others don't -- it often leads to re-negotiating the deposit procedure with publishers and upgrading the quality of deposit submissions. For publishers, this interaction helps them to redesign their publishing process according to higher quality standards.

In some cases a SIP may contain only metadata. This may be primary metadata, coming straight from the publisher or from identification agencies (national ISBN/ISSN agencies), or it may be a full bibliographic description coming from the library cataloguing system.

(8) Packaging & Delivery

This interface can request and accept a DIP (Dissemination Information Package) from the Access module of a DSEP. The DIP consists of the requested publication in one of the available formats with accompanying software and/or metadata needed to install and display it, to assess its authenticity or to reconstruct the original copy.

The interface takes care of all processes needed to unpack a DIP and to make it fit for use by the library visitor. Through this interface, deposited material can be made available, taking account of all kinds of access variables of the digital library environment, such as user authorisation, user access rights, publisher license access conditions and other access controls. Presently, for example, deposit license agreements with publishers only permit installation of publications onsite, on a library workstation, and access by registered library users.

This "post-processing" interface is needed because deposit libraries cannot anticipate

all access modes and future variables.

This interface also transfers, upon request, metadata from the DSEP through to other systems that need to process the data, either within the digital library system or external to it, such as systems from bibliographic utilities. Usually this concerns metadata uploads from DSEP to other systems. In some cases, it may also involve passing a whole publication through to a content indexing system, in order to generate, for example, a full-text index of the publication.

The main modules of a DSEP

The DSEP itself consists of six processing modules: the five OAIS modules, plus an additional module for preservation. The need for this additional module is clarified below.

1. Ingest

Ingest only accepts publications packaged as a SIP (Submission Information Package). Ingest unpacks and verifies the publication, collects, generates and re-distributes data to other processes. Routines include integrity check of the medium, of the file formats and of the logical document structure. The process identifies the informational contents, the primary metadata, special access controls to be placed on the contents, abstracts, full-text indexes and other additional data accompanying the publication, technical data for installation and de-installation. The different data are copied to and processed in different environments (for cataloguing, for access control, and for finding aids). In the process, the publication is installed and de-installed and its authenticity is established and recorded. Finally, Ingest prepares the publication for transfer to storage, as an AIP (Archival Information Package).

2. Archival Storage

Archival storage only accepts AIPs. This module consists of all procedures necessary for the secure storage of the electronic publication in the digital store, including storage management procedures, quality assurance, disaster recovery, etc. It also includes regular medium migration, in order to preserve the bit stream of a publication from decaying carriers.

3. Data-Management

Data-Management mainly stores and retrieves metadata. We distinguish between two types of metadata:

- metadata and technical data associated with the publication, such as bibliographic descriptions, access control information, (de-)installation data, authenticity and integrity control information, preservation data, etc.
- metadata associated with the administration of the DSEP, such as status report information, statistical data, etc.

The metadata associated with the publication may also be duplicated in other (external) systems. The cataloguing process, which creates a title-description of the electronic publication and also involves subject indexing, takes place in the cataloguing environment of the digital library system. It may re-use primary metadata provided by the publisher and return descriptive metadata to the DSEP system, through the Delivery and Capture interface.

4. Access

In the DSEP model the Access module is much more limited than in the OAIS model, because many related processes belong intrinsically to the digital library environment and not specifically to a DSEP, such as creating finding aids, registering library users, applying access controls, etc.

The DSEP access module takes care of retrieving an AIP and making it available in such a way that it is fit for use. This may entail extracting parts of the electronic publication, or adding a full-text index to it, or converting (parts) of the publication into appropriate formats for viewing, printing or downloading. It may involve providing a viewing configuration. It may even involve providing emulation software for displaying the publication. The resulting DIP (Dissemination Information Package) is then fed into the library access system.

5. Administration

The administration module is central to a DSEP. It regulates all the operations of the system and takes care of monitoring, quality control and auditing. It requests status reports from all processing modules and controls, regularly, if DSEP standards and policies set out by the deposit library management are applied throughout the system.

6. Preservation

The OAIS model does not explicitly include a preservation module. Medium migration (refreshing or copying a publication) is a preservation procedure that takes place in Archival Storage. It should be associated with storage because the stored bits need to be preserved. But archival storage does not have (and does not need to have) any knowledge of the content of a publication.

As formats become obsolete and the viewers needed to interpret and render these formats also become obsolete, it will be necessary to take measures to preserve the content of a publication and all related aspects such as data, layout, structure and functionality. To this end, several strategies may be followed, such as migration and emulation. In the OAIS model, digital migrations that require changes to the content are referred to as transformations. In all cases, transformation leads to a "new version" of the original publication. However, it is not clear where transformation processes take place in OAIS.

We have added a dedicated Preservation module to address this need. The module is configured according to the deposit library preservation policies. Both transformation and emulation approaches are worked out in some detail in the DSEP model. The resulting output is either a new version of a formerly deposited publication, in which case it is ingested anew in the system, or it is a set of specifications for building emulators that can render a whole generation of publications on a future (unknown) platform. In both cases, new preservation metadata will be generated and fed into Data-Management.

Data model for a DSEP

The data model for a DSEP is based on the OAIS information model. The deposit copy of an electronic publication is exchanged and managed within a DSEP as an OAIS information package (SIP, AIP and DIP). Such a package contains the following:

- The original bit stream of the digital publication
- Metadata

This may be primary metadata as provided by the publisher (title information, system requirements information, etc.) in the case of a SIP, or functional metadata necessary for specific functional entities (storage, preservation, access, etc.) in the case of an AIP or DIP.

- Software

This is the application software required to "render" the publication (viewer, browser, search and retrieval software, etc.), sometimes accompanying the publication in a SIP, and/or provided by the library in a DIP.

- Packaging information

This is data about the package being exchanged, such as package label, identifier, structure of content, etc.

However, it should be noted that the OAIS information objects are logical objects. In actual DSEP-implementations, the metadata, the software and the data bit stream need not be stored physically together in one AIP. In fact, it is proposed that, within a DSEP, all metadata is stored in Data-Management and not together with the data bit stream in Archival Storage. This is done because metadata updates will be more or less frequent, whilst the data bit stream of the publication content will not change over time. It is therefore not deemed sensible to store both types of data together in one physical container. All logical data entities that belong together and pertain to the same publication need to be linked together via interoperable identifier systems (identifier of the publication, of the information package, of the metadata records, etc.).

Metadata for preservation

The OAIS concepts of "Representation information" and "Preservation description information" allow for the correct interpretation of the data bit stream over an indefinite period of time. In a DSEP environment, the "Representation information" includes all technical characteristics of a publication, in particular:

- The **format(s)** of a publication, referring to the way in which the data is encoded (file formats, character encoding, etc.).
- The **navigational structure** of a publication, referring to directory/files structure of a publication (table of contents, numbered list of items, navigational structure with hyperlinks, etc.).
- The **application software** accompanying the publication, referring to the software that is required to "render" the publication (viewer, browser, search and retrieval software, etc.).
- The **system requirements**, referring to the hardware and systems software configurations that can run the publication. Examples of such system requirements are:
 - *for a publication on a CD-I*: Philips CDI 220 or the portable CDI 360 player with the Digital Video cartridge, TV with scart-connector, 6-inch colour LCD screen.
 - *for a publication on a CD-ROM*: IBM-compatible PC, 33 MHz 486 CPU or better, 4 MB internal memory, MS-Windows 3.1, 5 MB available on hard-disk for installation, CD-ROM player with adequate access time (less than 400 milliseconds), accelerated 256 colour VGA adapter compatible with MS-Windows (resolution 640x480), mouse or other pointing device compatible with MS-Windows.

In DSEP, the "Preservation description information" includes all recorded metadata giving information about the authenticity of a deposit copy and the preservation measures taken by the DSEP. Depending on the preservation strategy followed, both types of information need to change over time.

Management of change and versioning of AIPs are central to the migration strategy: when the original data bit stream is converted, the data formats, the rendering software, the system requirements and associated (de-)installation data, and the amount of information/functionality loss change and need to be recorded.

Assessing the digital original and authenticity control are central to the emulation strategy: choices need to be made as to what needs to be preserved of the original publication, what needs to be recreated (emulated) and what is an acceptable loss of authenticity. The parts that need to be emulated need to be specified in detail (metadata) in a high-level language and the user needs to be educated to "use" the

digital original -- as future generations will not know how to interact with obsolete IT-based end user environments.

The discussion of preservation strategies for deposit libraries is still continuing within NEDLIB. It is, however, clear that there is not one ideal strategy. Many aspects are involved, such as deposit conditions agreed upon with the publishers, cost aspects, future user access requirements, legal constraints, etc. NEDLIB partners are in agreement that we need more practical, hands-on, experience with different preservation approaches, in order to be able to evaluate their adequacy for deposit libraries.

Experimenting with emulation for preservation

In May 1999, the Koninklijke Bibliotheek and Jeff Rothenberg agreed upon a project proposal to perform emulation experiments for long-term preservation purposes. The overall purpose of the project is to test the viability of using hardware emulation as a means of preserving digital publications in a deposit library. The experiment will be designed to test and evaluate the hypothesis of Jeff Rothenberg, as publicised in 1995 in the *Scientific American* [ref. 3]. For the application area of deposit libraries, Rothenberg has formulated his hypothesis as follows:

"The original hardware environment (processor, display, peripherals, etc.) required to run the original software used to render digital publications can be cost-effectively described with sufficient accuracy to enable the creation of software emulators of that original hardware environment that can be executed by future host hardware environments, and using such emulators to run that original software on future hardware will render saved digital publications in ways that are sufficiently similar to their original renderings to qualify as preserving those publications in the manner required by a Deposit Library."

The project consists of re-iterations of basic experimental tasks such as designing the experiment, performing the experiment and evaluating the results of the experiment, in consecutive stages throughout 1999-2001. The first stage of the experiment, carried out during 1999, performs a "base-case" iteration of the experiment. This entails developing an initial experimental environment, with a selection of materials, a set of preservation criteria and well-defined procedures for testing and validation, and consequently performing "null" and off-the shelf emulation experiments. This first iteration serves to carry out a simplified end-to-end run of the experimental process, in order to calibrate it and to define in more detail the next iteration of the experiment. The initial iteration should result in the identification of relevant hardware aspects of platforms that need to be emulated to satisfy preservation criteria, as defined in this first stage. The second iteration, to be carried out in 2000, will aim to develop emulator-specifications for a representative range of original platforms and system configurations, as well as an experimental portable emulation environment that can interpret these specifications and host this environment on a reasonably wide range of different target platforms. An important requirement is that the specifications of the original hardware system necessary for emulation are available. The third, post-year 2000, iteration of the experiment will aim to refine and extend the emulator

specifications and emulator environment hosting techniques, to ensure their long-term viability.

The overall execution of the experiment will be done in a progressive way, step by step. This entails that, in the first stage, parts of the experiment will not be fully executed, but prototyped. In addition, the experimental variables, such as the samples of digital publications and the sets of preservation criteria, will differ with each iteration of the experiment. A full description of the experiment and of the process followed for preserving electronic publications in the emulation experiment will be made, in order to enable verification of the experiment.

The experiments will be verified at the Koninklijke Bibliotheek, using the testbed environment of the Deposit System for Electronic Publications (DSEP) developed in NEDLIB. The sample material to be used during the testbed experiments will be provided by the NEDLIB sponsoring publishers: Elsevier Science, Springer-Verlag and Kluwer Academic Publishers.

Finally, it is intended to incorporate the design of the emulation for preservation process into the OAIS Reference Model, as applied to Deposit Libraries by NEDLIB. The (meta)data elements required to preserve publications by means of emulation will be specified and represented in the NEDLIB data model.

References

[1] NEDLIB web-site: <http://www.konbib.nl/nedlib/>
[Back to the text](#)

[2] Referencing Model for an Open Archive Information System (OAIS),
<http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html>, White Book, Issue 5.0,
April 1999, Don Sawyer / NASA and Lou Reich / CSC.
[Back to the text](#)

[3] Rothenberg, Jeff. 1995. "Ensuring the Longevity of Digital Documents.",
Scientific American 272(1): 24-29
[Back to the text](#)

Copyright © 1999 Titia van der Werf-Davelaar

[Top](#) | [Contents](#)
[Search](#) | [Author Index](#) | [Title Index](#) | [Monthly Issues](#)
[Previous Story](#) | [Next Story](#)
[Home](#) | [E-mail the Editor](#)

[D-Lib Magazine Access Terms and Conditions](#)

[DOI](#): 10.1045/september99-vanderwerf