

NEDLIB Technical meeting paper

Document Number:

Title: Data model for a Deposit System for Electronic Publications

Date: 19-2-1999

Author: Titia van der Werf, Koninklijke Bibliotheek

=====

Foreword

This document builds on the concepts presented in the Reference Model for an Open Archival Information System (OAIS) [ref.1]. The draft-ISO technical recommendation is aimed at developing a consensus on what is required for an archive to provide permanent preservation of digital information. The ISO-work is seen as preliminary to adopting or developing specific standards needed to support archival type of services.

CONTENTS

- I. Data Model for a DSEP**
 - Content Information**
 - Characteristics of EPs and Starting Points**
 - Preservation Strategy**
 - DSEP Requirements for Representation Information**
 - Existing Practice**
 - Existing Standards**
 - Conclusions**
 - Preservation Description Information (PDI)**
 - DSEP Requirements for PDI**
 - Existing Practice**
 - Existing Standards**
 - Conclusions**
 - Packaging Information (PI)**
 - DSEP Requirements for PI**
 - Existing Practice**
 - Existing Standards**
 - Conclusions**
 - Package Descriptive Information (DI)**
 - DSEP Requirements for DI**
 - Existing Practice**
 - Existing Standards**
 - Conclusions**
- II. AIP, SIP and DIP: Package variants in a DSEP**
- III. Package transformations**
- IV. Data Flow**
- References**

Structure of this document

In this document the data model and data flow in a Deposit System for Electronic Publications (DSEP) is detailed. It is based on the OAIS-information model but it gives a more detailed interpretation of the model as applied to the handling of electronic publications. It incorporates the metadata requirements identified in the NEDLIB Functional Requirements document [ref.2].

Metadata requirements for preservation are dependent of the preservation strategy followed. In order to be able to apply the OAIS-concept of Representation Information in a DSEP environment, some characteristics of Electronic Publications (EP) are highlighted and some starting points concerning preservation strategies for a DSEP are formulated. These starting points align very well with Jeff Rothenberg's emulation scenario.

The document positions existing metadata models for bibliographic control, for resource discovery and for access control in the DSEP data model. It highlights which parts of the data model require further specification and which parts are already well developed in library systems.

I. Data Model for a DSEP

This section builds on section 4.2 of the OAIS Reference Model Document (OAIS).

An OAIS-Information Object is composed of a Data Object and Representation Information:

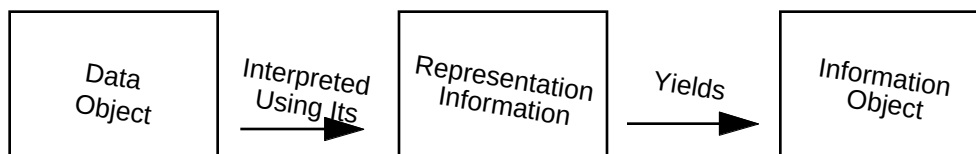


Fig. 1: Obtaining Information from Data [OAIS p.17 Figure 2-2.]

In the case of a DSEP, the Data Object is always a Digital Data Object. For preservation purposes a Digital Data Object is considered to be a meaningless bit-stream. It requires Representation Information that provides additional and sufficient information to be able to convert this bit-stream into meaningful information.

There are several types of information involved in the long-term preservation. The Content Information is the primary information of interest: it is the data object that needs to be preserved. Other types of information objects, related to the preservation of Content Information, are: Preservation Description Information (PDI), Packaging Information (PI) and Descriptive Information (DI). The relationship between these different information objects is illustrated in [Fig.2]. It is defined by the definition of an information package:

" An **Information Package (IP)** is a conceptual container of two types of information called **Content Information** and **Preservation Description Information (PDI)**. The Content Information and PDI are viewed as being encapsulated and identifiable by the **Packaging Information**. The resulting package is viewed as being discoverable by virtue of the **Descriptive Information**. " [OAIS, p.17]

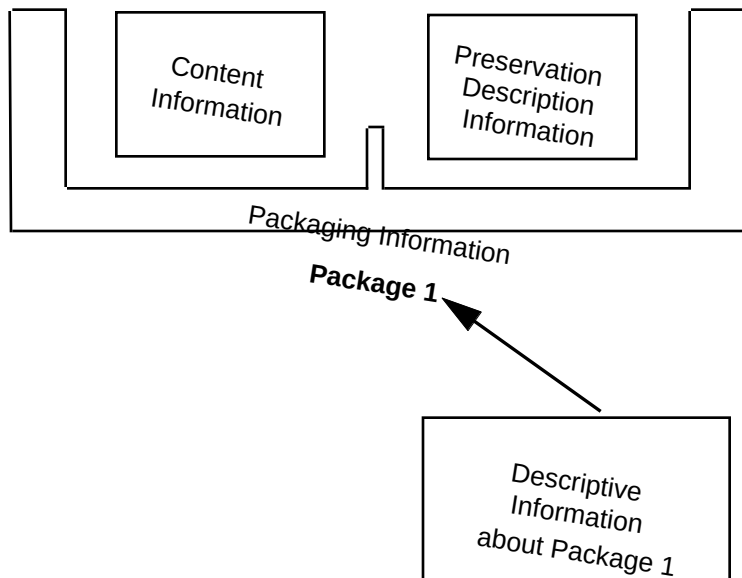


Figure 2. Information Package Concepts and Relationships [from OAIS Fig. 2-3, p.18]

Below the different OAIS-information types will be discussed in the DSEP context.

Content Information

In a DSEP the Content Information Object is composed of a Digital Data Object, which is the Electronic Publication (EP) itself, and the Representation Information that allows for the full interpretation of the Digital Data Object into meaningful information.

In order to define more in detail this Representation Information in the context of a DSEP, we need to look into the characteristics of EPs and we need to formulate some starting points concerning the preservation of EPs.

Characteristics of EPs and Starting Points for a DSEP

1. The deposit needs to preserve EPs in their original form

DSEP aims to keep electronic publications accessible in their original form, for the long-term. This means that a DSEP needs to keep more than only the informational content. A DSEP needs to keep also the presentation and functionality of the original publication and be able to recreate the "look and feel" of the publication any time in the future. If this starting point is accepted, Representation Information needs to preserve (detailed) information about the original software and hardware configuration accompanying the EP, in order to be able to recreate this environment. If this starting point is not accepted, transformation of the original electronic publication is permitted with inevitable loss of information. Instead of Representation Information there would be a requirement for recording Loss of Information. Metadata requirements for preservation are dependent of the preservation strategy followed.

2. The deposit needs not preserve the original processing software

Providing access to deposited EP does not require support of the original processing software. Users do not need to process the information content of EPs with the original software. Access requirements only need to support those facilities that were offered at the time of publication, such as viewing, printing and downloading. If this starting point is accepted, Representation Information need NOT preserve the software context needed to manipulate EPs, such as text-processors, spreadsheet software, etc. Only the software for viewing the EP is required.

3. Relative longevity of hardware configurations for EPs

Another starting point for a DSEP is that publishers tend to market their products on media for the masses. They tend to distribute their products via channels and in formats that are readily available on the consumer market, such as CD-ROMs. This means that the hardware and software configurations for EPs will be quite "standard" during a period of time. These standard configurations will support relatively many publications published in a given period of time. In addition, the software configurations will tend to evolve faster than the hardware configurations. This in turn means that for a given hardware configuration a greater quantity of EPs is catered for compared to the amount of EPs supported by a given software application [see Fig.3]. The preservation strategy of a DSEP should take this layered model of the longevity of EPs, software and hardware, into account.

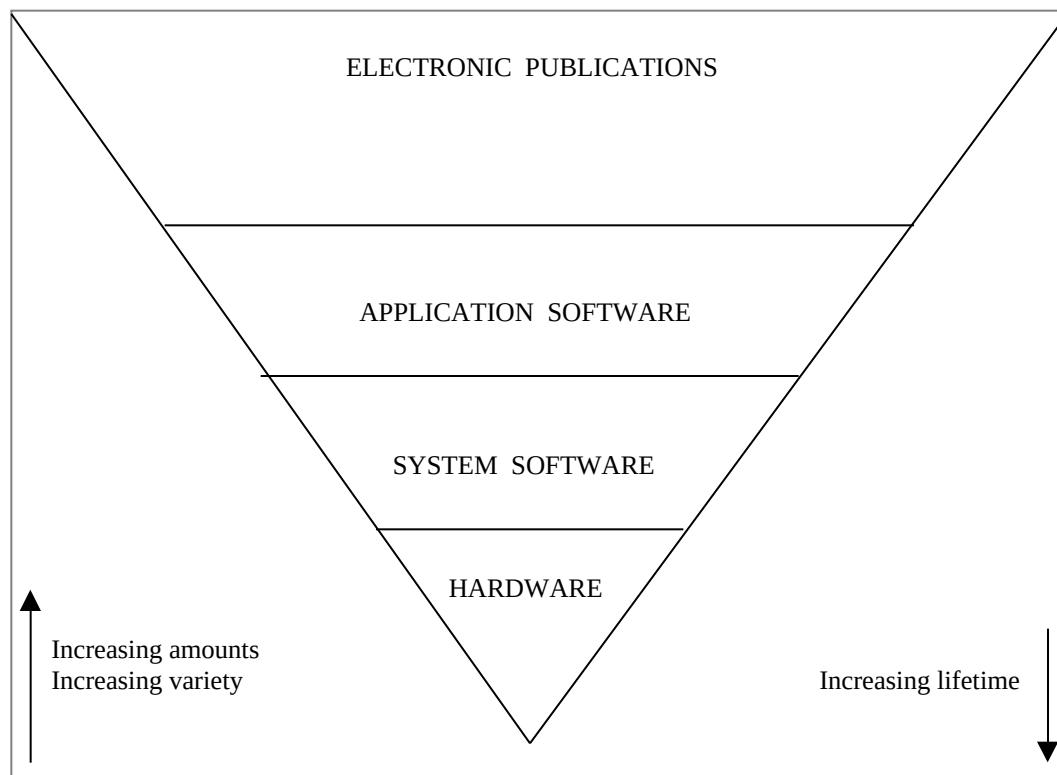


Fig. 3. Layered longevity model: longevity of EPs and their required hardware and software configurations [Adaptation of Jeff Rothenberg illustration]

Preservation strategy for a DSEP

From the starting points listed above an appropriate preservation strategy for a DSEP can be formulated as follows:

1. EP + Representation Information (Software + Hardware)

Ideally an EP should be kept together with the hardware and software configuration that supports viewing the publication.

2. Hardware Emulation

As the hardware configuration becomes obsolete and is replaced by new hardware (components), the software cannot execute properly and the publication cannot be viewed any longer. To recreate the appropriate executable environment for the software, the hardware (components) can be emulated.

With emulation of the hardware (components), the preserved software configuration can, in theory at least, keep functioning even as technology evolves and new software/hardware platforms replace older ones. For preservation purposes it is therefore more efficient to emulate the hardware configuration only and to preserve the software together with the EPs.

With this strategy we align ourselves with the proposal made by Jeff Rothenberg, i.e. “to save records (data, documents, etc.) along with the application software that created them, as well as the supporting system software needed to run this application software and a description of the required hardware environment needed to run this entire suit.” (ref. p.8).

3. Software library

The hardware emulation strategy is particularly relevant in a DSEP environment. As stated earlier, the software that needs to be preserve with the EP is not unique for that publication but caters for a whole generation of publications. Therefore a DSEP can best keep a software library that can be referred to, instead of keeping a copy of the same software with each publication.

It should be noted that this strategy is not necessarily well suited for archives where software is more customised to the particular needs of government organisations and therefore less consumer market oriented. In addition for archives the preservation of the whole software processing environment may be required, not just viewing, printing and downloading.

It should also be noted that the strategy outlined above is only very generally applicable and needs some refinement. There may be different ways to apply it, according to publication genres:

- Marked-up files that can easily be interpreted only need to be accompanied by a specification of the mark-up language. Developing an interpreter for these documents may be considerably more economical than building hardware emulators to run the old viewer software.
- Some hardware/software configurations may not permit emulation of hardware only (direct register addressing by the software etc...)
- Etc.

Finally, if references to a software library are made, the software library should be preserved to. It may be interesting to investigate the feasibility of a common software library for several DSEPs (within a common language region for example). Keeping a software library separately may have advantages over keeping a copy of the software with each document. Keeping the hardware descriptions/specifications together with the software library may be more efficient than keeping them with the publications.

However, in general we can formulate the requirements for Representation Information in a DSEP as follows:

DSEP Requirements for Representation Information of Content Information:

The Representation Information for an EP should include (information about) the hardware and software configuration that supports viewing of the publication, in particular:

1. (a reference to) the software required to view the EP.
2. the description of the required hardware in order to be able to emulate it.

The Representation Information for EPs will provide at least the following:

- the necessary information for the operating system to access the bit-stream on the media layer of the EP or of archival storage system (CD-ROM, disk, tape, etc.).
Note: an initial media migration might take place to move the EP from the publisher's distribution medium to the archival storage medium. This type of information includes filenames and directory structures (location address within the associated bit sequence) that can be recognised and operated by a given operating system.
- (reference to) the operating system needed to interface with the programs that can render the EP for viewing.
- (reference to) the programs necessary to view the EP
- (reference to) the description of the hardware configuration required to run the programs.

Existing practice: what are deposit libraries using to record Representation Information?

Up to now libraries record software and hardware (system) requirements as follows:

- in special fields of the bibliographic description format
- the description of system requirements is textual and human-readable
- the requirements description is based on information supplied by the publisher
- through installation and de-installation of the software. In most cases this is done manually.

In some cases the system requirements and setup configurations may also be encoded in machine-readable form to enable automated processing of EPs (MMB-system).

Existing Standards:

ISBD ?

(From NEDLIB standards and conventions document)

USMARC Format Standard:

- 008 control field, codes for computer files
- 256 computer file characteristics
- 340 physical medium (subfield i)
- 516 type of computer file or data note

- 535 location of originals/duplicates note
- 538 system details note
- 562 copy and version identification note
- 753 systems details access to computer files
- 856 electronic location and access

Conclusion

Existing practice for recording Representation Information in a DSEP is not sufficient. A DSEP requires the machine-readable encoding of this information in more detail and in a specified format. There should be an agreed standard for referencing to information outside the information package.

Preservation Description Information

The preservation description information applies to the Content Information and is needed to make this information understandable when the knowledge base of the customer no longer contains the necessary information. This type of information provides for:

- **Provenance and Context.** These two OAIS-categories of PDI provide for the **contextual information** necessary for a correct interpretation of the Content Information: how the Content Information relates to other information objects in/outside the package (Relations such as IsPartOf, IsVersionOf, HasSource, etc.), how the Content Information came about (history/provenance)
- **Reference and Fixity.** These two OAIS-categories of PDI provide for the required **identification and authentication information**: such as info on the identification (unique number/reference/address), authenticity (certificate, digital signature) and integrity (checksum) of the Content Information.

Note: The OAIS-document gives examples of these categories for bibliographic information (Table 4-1, p. 56):

Content Information Type	Reference	Provenance	Context	Fixity
Space Science Data	Object Identifier Journal Reference Mission, instrument, and title attribute set	Instrument Description Processing History Sensor Description Instrument Instrument mode Processing history Decommuration map Software Interface Spec.	Calibration history Related data sets Mission Funding history	CRC Checksum Reed-Solomon coding
Bibliographic Information	ISBN Title Author	Printing history Copyright Position in series References	Related References Dewey Decimal System Publishing Date Publisher	Author Digital signature Cover
Software Package	Name Author/Originator Version number Serial Number	Revision History License holder Registration Copyright	Help file User Guide Related Software Language	Certificate Checksum Encryption CRC

The examples given for bibliographic information are not very appropriate. Those given for "software package" probably give a better idea of the type of metadata required for EPs.

- electronic publications in the case of a DSEP
- archival records in the case of an archive

For each category of PDI, examples of relevant metadata elements are given:

Content Information Type	Reference	Provenance	Context	Fixity
Electronic Publications	- unique identifier of the EP (ISBN, DOI, ISSN, etc.) - edition - version number - issue number - title - author	- Edition / version history - Publisher - Publishing date and place - Copyright	- helpfiles, user manual, - related software - references (IsDerivedFrom, IsFormatOf) - language	- digital signature - copyright registration certificate - checksum - encryption
Archival Records				

DSEP Requirements for Preservation Description Information

The *contextual information* (Provenance and Context) requirements for EPs are not different from those for printed material. Existing bibliographic description standards and practice provide for all required metadata elements.

The *identification and authentication information* (Reference and Fixity) for EPs is not so straightforward.

Identification

Publishers of EPs are developing new identification systems because the traditional schemes for printed matter do not meet requirements of electronic publishing. The publishing industry is mainly investing in the DOI-system which should provide a unique identifier for digital objects together with the necessary metadata for the identification process.

Authentication

Authentication information for digital material is different from that for printed matter. Here too digital encrypting standards are meeting new requirements, such as digital signatures for authentication and hash techniques for file integrity checking. Authentication in the sense that you want to make sure you have an authentic copy of an EP is possible in a DSEP as long as copying of the EP within DSEP, such as refreshing and replicating for storage and preservation purposes, does not affect the EP as such. In this sense each copy is still authentic. A DSEP should be able to furnish such a guarantee of authenticity. However when an authentic copy has to be transformed in the DSEP into a new archival version of the EP, affecting changes in the Content Information, the new version is no longer authentic. It will also require a new integrity checksum, because the original one will not apply any longer.

Representation Information for PDI

The PDI-object will also require representation information. PDI is encapsulated together with Content Information and both need representation information to be interpretable. For the PDI, representation information needs to provide for a listing of the standards used to encode the PDI: the metadata standards, the identification

standards and the authentication standards. Not all of this information is necessarily text-based (for example the M5D hash).

Existing practice

Provenance and Context are described and supported by existing bibliographic standards.

Identification and authentication of EPs are not supported adequately by the publisher community yet. Verification procedures for EPs are not yet in place at the deposit libraries. Integrity checking does occur when the EP is accepted and registered. Integrity checking information is also recorded in some instances for file comparison at later stages (for example web harvesters make use of M5D). Authentication information is usually not recorded.

Both identification and authentication have been research areas of BIBLINK.

Practical recommendations from BIBLINK are the following:

- use the unique identifiers provided by publishers to their EPs
- if a publisher fails to provide an identifier with the EP, the NBA should allocate an identifier to the EP
- ask the publisher to provide an M5hash of the EP for integrity checking.

Existing Standards

(From NEDLIB standards and conventions document)

Provenance and Context: ISBD / MARC standard

Identification: existing and emerging standards (ISBN, ISSN, SICI, DOI, ...)

Authentication:

- integrity check: M5Dhash
- authentication: certificate information from publisher (digital signature)?

Conclusions

Provenance and Context are described and supported by existing bibliographic standards.

Identification and authentication of EPs are not supported adequately by the publisher community yet. Project BIBLINK is addressing these issues. NEDLIB should follow BIBLINK recommendations.

Concerning authentication BIBLINK has only addressed the integrity issue.

In NEDLIB metadata should support authentication information of archival versions of EPs: archival versions can either be authentic or not. This should be recorded as part of the PDI.

In NEDLIB Representation Information for the PDI should also be supported.

Packaging Information

The packaging information binds the [Content Information + Representation Information] and the [Preservation Description Information + Representation Information] together. This is done through identification/location information of the Content Information Object and of the Preservation Description Information Object (IdPI => IdCI and IdPDI). It's a local access resolver for the information contained in the package. It also contains information such as package version number, package type, etc.

DSEP Requirements for Packaging Information

The identification requirements are internal identification requirements to a DSEP:

- IdPI: identifier of the Packaging Information Object
- IdCI: identifier of the Content Information Object
- IdPDI: identifier of the Preservation Description Information Object

The location requirements are also local to a DSEP. The location of Information Objects should be standardised throughout the DSEP.

The information should clarify the nature of the Information Package: SIP, AIP or DIP.

The information should carry versioning information of the package. If for example through migration repackaging is required, the packaging information is changed and location information of Information Object may have changed. Versioning of packaging information is therefore required.

Representation information for PI

The PI-object will also require representation information. This representation information should be a listing of the standards used for encoding the identification, location, typology and versioning information.

Existing practice

MMB practice?

IBM-DL?

Existing Standards

?

Conclusions

A DSEP requires minimal packaging information, including the local identification of Information Objects, a standard method for locating Information Objects, versioning of packaging information and information on the type of Information Package. Representation Information is also required.

Descriptive Information

The descriptive information is that information which is used to discover which package has the Content Information of interest (Description of Content => IdPI).

Descriptive information is derived from PDI and Content Information.

DSEP Requirements for Descriptive Information

This information is necessary to find and identify the Content Information. It is descriptive and identification metadata. The descriptive metadata requirements for EPs are not basically different from those for printed matter.

The identification requirements for DI should be the same as those described for PDI.

Both DI and PDI should identify the same Content Information with the same identifier, i.e. the identifier of the EP, allocated by the publisher.

The DI should also lead to the location of the requested Content Information: this means DI should contain identification/location information of the Archival Information Package (IdPI).

This resolution (IdCI => IdPI) should be supported within the finding Aid environment. This entails local resolution within the digital library (note: this is not the same as local resolution within DSEP! DSEP is part of a bigger context: the digital library. The finding aids and the resolution system are outside of DSEP.).

There may also be a requirement for the (global) last-resort resolution of identifiers. This is the case when a publisher assigns, for example, a DOI to an EP, deposits a copy of the EP in a DSEP and decides after the commercial life-time of the EP to stop electronic delivery of copies of the EP. As a consequence the (global) DOI-resolution-system cannot resolve the DOI. It re-directs the DOI to the (global) networked deposit libraries resolution system (for example, by re-encoding the URN:DOI-string into a URN:NBN:DOI-string).

Representation Information for DI

Descriptive information lies outside of the Information Package. In the digital library environment it is part of information maintained in operational database systems. These database systems are continuously updated and important (structural) data changes involve migration of the whole database system to an upgraded version. As long as the database systems are operational, the data maintained in them is accessible and does not require representation information.

Existing practice

EPs are described according to existing bibliographic practice. Subject indexing is performed. Content-indexing too.

The identifier of the EP is also part of the bibliographic description.

Location information is recorded in the descriptive record.

Existing Standards

Bibliographic Description Standard (ISBD)

Classification Standards (LCSH, Dewey, UDC, etc...)

Identification standards: ISBN, ISSN, SICI, DOI, ...

Location: URL, deposit shelf number

Conclusions

In the digital library environment there are existing finding aids (catalogue, full-text indexes, abstracts, ..) and standards that can support the discovery of EPs archived in a DSEP.

There is no need to define any additional ones.

There is a requirement for local resolution of identifiers to the location of the archival information package inside the DSEP.

There may be a requirement for the (global) last-resort resolution of identifiers.

There is no need for representation information of DI.

Package variants

In the OAIS-document it is stated that it is "necessary to distinguish between an Information Package that is preserved by an OAIS and the Information Packages that are submitted to, and disseminated from, an OAIS." (OAIS section 2.2.3, p. 19)

These variant packages are needed:

- to reflect the reality that some submissions to an OAIS will have insufficient Representation Information or PDI to meet final OAIS preservation requirements
- because information submissions may be organized very differently from the way the OAIS organizes the information it is preserving
- because the OAIS may provide information to Consumers that does not include all the Representation Information or all the PDI with the associated Content Information being disseminated.

The variants are referred to as the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP).

In a DSEP environment the (minimum) requirements for each variant package need to be specified. This will be done in NEDLIB by WP2 (D2.1.1), WP4 (D4.1.1) and WP3 (D3.1.1) respectively. This document will be the basis for this work to be carried out.

The **Submission Information Package (SIP)** is the information package, containing the EP, that is received by the DSEP from a publisher. Its form and detailed content is typically negotiated between the Publisher and the DSEP (see also Deposit Guidelines).

Most SIPs will have:

- Content Information, with some Representation Information
- some PDI (see BIBLINK core)
- Packaging Information will always be present in some form

It may require several SIPs to provide a complete set of Content Information and associated PDI to form an AIP.

A single SIP may contain information that is to be included in several AIPs.

Within the DSEP one or more SIPs are transformed into one or more **Archival Information Packages (AIP)** for preservation.

The AIP has:

- Content Information, with complete Representation Information
- a complete set of PDI for the associated Content Information
- Packaging Information of the AIP will conform to DSEP internal standards.

In response to a request, the DSEP provides all or a part of an AIP to a Library User in the form of a **Dissemination Information Package (DIP)**. The DIP may also include collections of AIPs.

The DIP will have:

- Content Information (with/without representation information depending if for example emulation is done outside or within the DSEP environment)
- Some PDI
- Packaging Information will always be present in some form so that the Library User can clearly distinguish the information requested. The Packaging Information may take several forms depending on the dissemination media and User requirements.

Package transformation

Data Flow

References

- [1] Referencing Model for an Open Archive Information System (OAIS), White Book, Issue 4, Don Sawyer / NASA and Lou Reich / CSC
(URL: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html)
- [2] Jose Borbinha, Fernando Cardoso, "Functional Specification for DSEP", NEDLIB project Document 6048/DEL-34, 15-10-1998
- [3] Jeff Rothenberg, "Ensuring the Longevity of Digital Information", 27-03-1998
- [4] Jeff Rothenberg, "Metadata to support data quality and longevity", 10-08-1998
- [5] Michael Day, "Metadata for Preservation", CEDARS Project Document AIW01, 21-11-1998
- [6] PANDORA Logical Data Model, version 2, 10-11-1997
- [7] Godfrey Rust, "Metadata: the right approach. An integrated model for descriptive and rights metadata in e-commerce.", D-Lib magazine, July/August 1998
- [8] "Functional Requirements for Bibliographic Records", Final Report of the IFLA Study Group, UBCIM Publications - New series vol. 19. München 1998
- [9] Norman Paskin, "The digital object identifier initiative: metadata implications", DOI discussion paper number 2, 25-11-1998