Quality Requirements of Migration Metadata in Long-term Digital Preservation Systems

Feng Luan[1], Thomas Mestl[2], Mads Nygård[1],

[1]Department of Computer and Information Science
Norwegian University of Science and Technology
Email: {luan,mads}@idi.ntnu.no

[2]Research and Innovation, Det Norske Veritas
N-1322 Høvik, Norway
Email: thomas.mestl@dnv.com

Abstract. Migration is the most common preservation strategy in long-term digital preservation systems. It relays old digital objects from one technique to another. A prerequisite for a successful migration is the availability of migration metadata, which provide enough background information to preserved digital objects. Without the migration metadata, the migration may not be possible, but also the digital object's consistency may be violated. It is therefore recognized that the migration metadata are essential but surprisingly no requirements on these migration metadata seem to be available. In this paper, quality requirements of such migration metadata are derived from common preservation metadata schemas. The completeness and the usefulness of these quality requirements are validated by a case study. The final results show that the quality requirements can actually improve the workflow of a migration procedure. In addition, they can be used to improve metadata schemas and thereby decrease the risks in future migrations. Finally, six improvement suggestions for preservation systems are derived from our analysis of the quality requirements.

Keywords: requirements, metadata, migration, preservation systems

1 Introduction

Migration may be considered as the most important preservation strategy [1, 2]. Waters & Garrett [3] defined it as "a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequence generation". Any migration is supposed to mitigate some threats to a preservation system, such as obsolescence of hardware/software, failures of hardware/software, bad performance, incompatible hardware/software, failures of a security mechanism, obsolescence of a reference technique, and changes in organizational requirements.

The Reference Model for an Open Archival Information System (the OAIS) [4] as well as Wheatley [5] addresses various migration types. However, no matter what migration type is chosen, a successful migration relies on a wealth of information stored as migration metadata. These migration metadata provide background about previous situations. If no migration metadata exist, any digital object may not be accessible (e.g. no device and its driver information) or may not be rendered (e.g. no information about its file format or encryption algorithm). This would terminate all migration solutions.

Having no migration metadata may also threaten the consistency of the digital object. For example the value of a digital signature depends on the signature algorithm. In addition, different migration solutions throughout the preservation time may affect the trustworthiness of a digital object as curators may use different approaches. Therefore, the importance of migration metadata is commonly recognized, but interestingly no detailed lists of quality requirements to these migration metadata have been given.

In the following, we will propose eleven quality requirements that migration metadata should fulfill, so that any migration can be executed in future. These quality requirements are derived from preservation metadata considered centrally in three preservation metadata schemas listed in Section 2. The quality requirements can be sorted into five groups, i.e., hardware, application, specification, object, and policy. These quality requirements are further analyzed in a format migration design workflow with respect to the input/output information for each task of the workflow. Our analysis indicates the necessity and the usefulness of the quality requirements. The quality requirements may therefore be used as guidelines when designing migration metadata and checklists for examining an existing metadata schema. Finally, we discuss the three metadata schemas in terms of the quality requirements and give six suggestions to long-term preservation systems.

The next section describes the current situation of the preservation metadata. In Section 3, the quality requirements are proposed, which are further exemplified by the cast study in Section 4. The paper ends with a comparison and six suggestions to a long-term preservation system.

2 Related works on preservation metadata

Metadata are supposed to provide background information about a physical or digital object. Over the last 40-years, the concept of metadata has evolved into a series of powerful tools such as Dublin Core set [6], XML schema [7, 8], Extensible Stylesheet Language (XSL) [9], and METS [10].

A sub-set of these metadata, called preservation metadata, ensures that any preserved digital object can be accessed and used over a long time frame. Lavoie and Gartner [11] defined five requirements to the preservation metadata: *Provenance* recording the change history, *Authenticity* validating integrity, *Technical Environment* describing technical requirements, *Rights Management* defining permitted operations.

Detailed information about preservation metadata can be found in the Archive Information Package (AIP) of the OAIS, containing the Content Information and the Preservation Description Information. The Content Information contains the Data Object and the Representation Information. The Data Object is the preserved digital object, whilst the Representation Information is necessary information for representing this digital object. The Preservation Description Information provides necessary descriptive information including the Reference Information, the Provenance Information, the Context Information, and the Fixity Information. The Reference Information set identifier mechanisms used for internal and external access purpose. The Provenance Information tells the history and modifications of the digital object. The Context Information describes the relations to other objects, and the Fixity Information provides the validation mechanism for the integrity of the object.

Although the OAIS did not define any schema for the preservation metadata, other initiatives derived such preservation metadata schemas, such as the Preservation Metadata for OAIS (the OAIS Metadata) [12], the Preservation Metadata - National Library of New Zealand (the NLNZ Metadata) [13], and the Preservation Metadata: Implementation Strategies (the PREMIS) [14]. The OAIS Metadata strictly complies with the AIP structure mentioned in the above paragraph. It further aggregates four existing preservation metadata schemas into one.

The NLNZ Metadata is constructed by four entities: File, Object, Process and Metadata Modification. Any archived information is mapped to the File. The File provides the profile of the archived information. Each File may contain several components represented as the Objects. Each Object is not only associated with the Processes describing the execution of preservation processes but also associated with the Metadata Modifications providing the change history of the preservation metadata.

The PREMIS defines five semantic unites, i.e., Intellectual Entity, Object, Agent, Right, and Events. The Intellectual Entity refers to an intellectual unit, for example a book, an image, music, a movie, or a database. The Object is a digitized form of the Intellectual Entity. The Agent is a person, an organization, or applications associated with one or more Events to the Object. The Right specifies the assertion of one or more rights and permissions to the Object or the Agent. The Event describes history of actions performed on the Object. The PREMIS defines data dictionary for other unites besides the Intellectual Entity since the Intellectual Entity is assumed to be covered by other descriptive data, such as Dublin Core Set.

3 Quality Requirements of Migration Metadata

Migration metadata is a sub-class of the preservation metadata providing background information about earlier states of the preserved digital objects. The function of the migration metadata is to assure that forthcoming migration procedures can be successfully executed. In this section, the migration procedures and the preservation environment will be analyzed with respect to what documentation is necessary, i.e., quality requirements of migration metadata (See Table 1). The quality requirements are derived from common metadata schemas such as the OAIS Metadata, the NLNZ Metadata, and the PREMIS.

Table 1. Quality Requirements of Migration Metadata

Category	Req.	Description
Hardware	*R1*	Metadata for documenting a storage medium and its driver must be preserved as long as the store medium is used in the preservation system.
	R2	Metadata for documenting a hardware environment must be preserved as long as one of the migration procedures depends on this hardware environment.
Application	*R3*	Metadata for documenting a transfer application must be preserved as long as the associated file system is used in the storage medium of R1.
	R4	Metadata for documenting an interpretation application must be preserved as long as its related format is used in the preservation system.
Specification	*R5*	Metadata for documenting any format technique that is used in the preservation system must be preserved.
	R6	Metadata for documenting any reference technique that is used in the preservation system must be preserved.
	R7	Metadata for documenting any security technique that is used in the preservation system must be preserved.
Object	*R8*	Metadata for documenting characteristics of any digital object must be preserved.
	R9	Metadata for documenting the change history of any migrated object must be preserved.
Policy	*R10*	Metadata for documenting the usage rights of any digital object must be preserved.
	R11	Metadata for documenting the retention policies of any digital object must be preserved.

3.1 Hardware

The *Hardware* refers to all hardware components that are necessary for retrieving the bits of the digital object for running any required software. Without metadata about the storage hardware, users may not access the digital object on the storage medium. Any migration attempt is stopped. The profile and the usage document of the storage and its driver should be preserved (see *R1*).

Further, if the digital object can only be read with certain software, the hardware environment information that supports the software should be preserved (see *R2*).

3.2 Application

The *Application* refers to software that is able to manipulate a digital object. In general, three types of software may be involved in a migration procedure. The first two types should be identified with respect to the migration metadata. The last type is decided when designing a migration plan. The first type is a transfer application that reads, transfers, and writes a digital object between two storage systems. The transfer application enables the migration procedure to access the digital object and save the migrated object. The transfer application is often included in the operating system (OS) as a basic function. To find and use this software, the profile, the installation document, and the usage document should be held (see *R3*).

The second type is an interpretation application that decodes a technique used in a digital object, such as a format technique, a reference technique, or a security technique. In the migration procedure, the interpretation application is often used to compare the digital object before and after the migration. Like *R3*, the profile, the installation document, and the usage document should be held (see *R4*).

The third type is a conversion application that converts any technique used by a digital object, such as a format technique, a reference technique, or a security technique. This conversion happens often as a by-product in an update of the preservation system. However, documents about such conversion applications are not mandatory to preserve since the conversion application should be assessed and compared each time.

3.3 Specification

The *Specification* refers to the documentation that defines the syntax, the protocol, and the grammar of a technique. The specification makes it possible to realize any old technique that is in use. Three types of techniques are often used in a preservation system. The first type is the format technique. The OAIS [4] defines that "a format is reversible, byte-serialized encoding of an abstract information model, which is itself a formal expression of exchangeable knowledge". If no format specification is documented, no one can develop an interpretation application or a conversion application to the format. Therefore, the format profile and specification should be preserved (see *R5*).

The second type is the reference technique used to connect two related digital objects. Two methods are usually used for referencing. One way is via a *link* that uses a hyperlink [15] to connect two points or regions of digital objects. The other way is via an *identifier* that is a unique name given to each digital object, for example Uniform Resource Identifier (URI) [16]. A migration procedure may change a digital object's location and format, or the updated preservation system may abandon the old reference technique. Thereby, the migration procedure probably breaks a reference. In order to recreate the broken reference, the old profile and the specification of the reference technique should be preserved (see *R6*).

The third type is the security technique. The security technique does not only restrict access and encrypt the digital object but also assures the integrity and authenticity of the digital object. Different migration procedures may require different security techniques. For example, when transferring bits between storage media, the checksum metadata is used to prove no bit loss; when converting a format of an encrypted digital object, the digital object should be decrypted first. In order to keep the security of a digital object, the profile and the specification of a security technique should be preserved (see *R7*).

3.4 Object

The *Object* refers to metadata related to the characteristics and the change history of a digital object. The characteristics refer to significant properties in content, rendering, structure, and behaviors [17]. In a migration, the function of the characteristics is to determine whether a migrated object is equivalent to the original (see *R8*).

The change history describes the previous activities and provenance to the digital object. Some migration procedures may modify a digital object. The difference between the current and the original version should be annotated and documented especially if the original version is not available. In order to keep provenance and authenticity, the change history should be preserved (see *R9*).

3.5 Policy

The *Policy* refers to laws, rights, and rules that could potentially prohibit migration. Two classes of policies are often used as a filter in migration procedures. The first class addresses the usage rights of a digital object including intellectual property rights (IPRs) and government laws. For example, the IPRs restrict operations on digital objects and the government laws specify the minimal lifetime of a document. Therefore, the usage rights that prove the legality of a migration procedure should be preserved (see *R10*).

The second class is the retention policy specifying a preservation level and the relevant rules for comparing migration solutions. In this way, curators can always use the same way to assess whether the migrated object deviates from the original. Therefore, the retention policy that keeps the migration consistency should be preserved (see *R11*).

4 Case Study: Designing a Migration Plan

Strodl et al. [18] designed a workflow for selecting format migration solutions so that digital objects can be safely preserved in a long-term preservation system. The usefulness of such a workflow is well recognized by many national libraries. In the following the quality requirements will be applied on this workflow demonstrating its impact and usefullness, see Figure 1.

The original workflow has three phases with a total of 11 tasks. The first phase, called *define requirements,* has three tasks. It starts with defining the migration scenario (i.e., Task 1), which includes current polices and usage rights (*R10*). In Task 2, curators apply the current policies and the *R10* to select sample files that will be the test files for evaluating the migration solutions. In Task 3, curators should also identify all the characteristics (*R8*) of digital objects and the conversion application's characteristics (*R11*).

The second phase (i.e., *evaluate alternatives*) provides a preliminarily evaluation of the migration solutions, including smaller test migrations. In Task 4, curators identify the important information about the old hardware platform (*R2*), the old storage system (*R1*), and all the old specifications (*R5*-*R7*). Then, based on this aggregated information, format migration solutions, including conversion applications and new file formats, can be chosen. Task 5 is an assessment of the conversion applications with respect to the current policies and the application's characteristics, such as performance, reliability, cost and desired functionality. The current policies have to be refined until a conversion application is considered satisfactory and then a migration plan can be designed in detail. In Task 6, the old and new technical situation should be stored as background information for any future migrations, i.e., the hardware (*R1*-*R2*), the applications (*R3*-*R4*), the specifications (*R5*-*R7*), and the mechanism for capturing the migration results that are changes on the digital objects (*R9*). Task 7 executes a trial format migration using the sample files and the experiment plan. The migrated files will be evaluated in Task 8. The original file and the migrated file are compared and interpreted in terms of both the object's characteristics (*R8*) and the application's characteristics (*R11*).

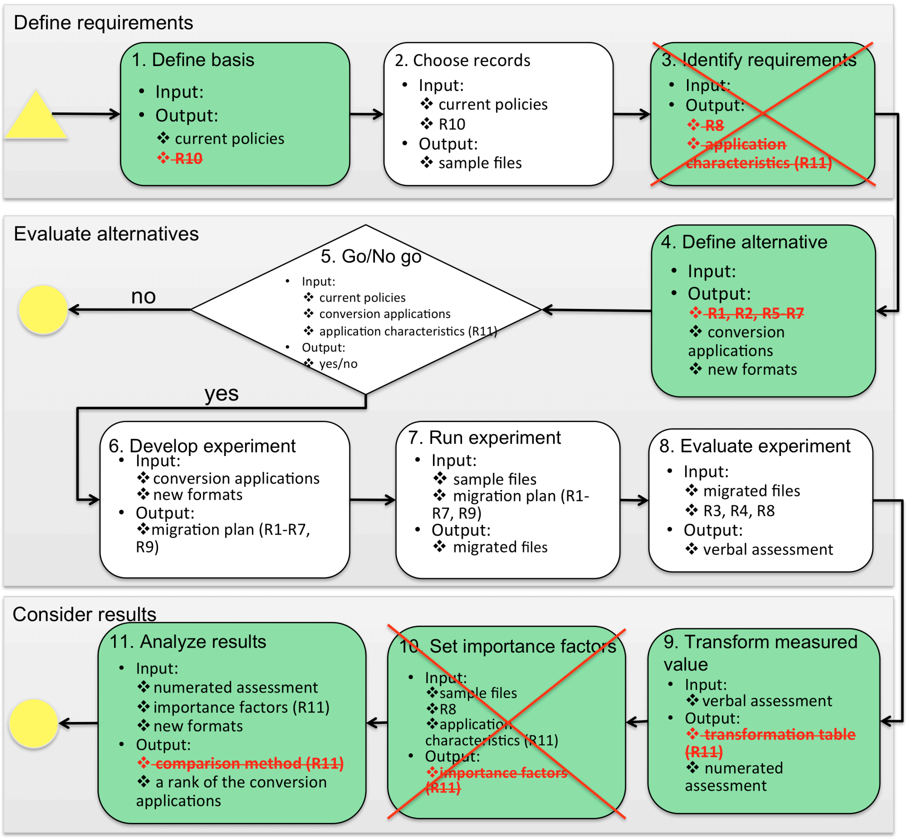

Fig. 1. A workflow for selecting a format migration solution. The workflow starts at the triangle and terminates at the circle where rounded boxes denote tasks having input/output information. Tasks that are affected by our quality requirements are marked as dark (or green) boxes with red color crosses or strike-throughs indicating their effect.

In the last phase (i.e., *consider results*) the migration solutions are ranked. The verbal assessments from Task 8 is converted into a scoring, i.e., a transformation table (*R11*) from verbal to scores. In Task 9, curators set importance factors to the object's and application's characteristics, *R8* and *R11* respectively. Before ranking the migration solutions in Task 11, a comparison method (*R11*) should be agreed on allowing to compare the final scores of an alternative formats. The format with the highest score may be chosen as the next format for the long-term preservation system.

By analyzing the input/output information in the original workflow, it becomes apparent that the workflow can be optimized if previous migration metadata are available. The improvements shown in Figure 1 are listed as follows:

- In Task 1: If information about the usage rights is available, the curators do not need to address this aspect anymore. Therefore, the curators do not need to contact the owners of the digital objects to obtain the usage rights (*R10*) again.
- In Task 3: If the characteristics of digital objects (*R8*) and the characteristics of a conversion application (*R11*) are available, curators do not need to discuss and design those characteristics again. The curators can obliterate Task 3 and go directly from Task 2 to Task 4.
- In Task 4: If the metadata about the previous hardware (*R1*-*R2*) and specifications (*R5*-*R7*) are available, curators can directly use this information to find suitable migration solutions. The curators do not have to look for the technical background again, i.e., *R1*, *R2*, and *R5*-*R7*.
- In Task 9: If the numeration rules are available, curators are able to directly transform the verbal assessment to number scoring. They do not need to consider the classification of scales anymore. Therefore, in Task 9 the discussion about the transformation table is unnecessary.
- In Task 10: If the importance factors of the characteristics defined in *R11* are available, then the output of this task is directly given completely. Task 10 can be omitted.
- In Task 11: If the description of the comparison method is available, curators can directly use it to compare various migration solutions. Therefore, the search for suitable comparison methods (*R11*) can be dropped in Task 11.

5 Discussion and Suggestions

Many applications have been implemented for extracting preservation metadata, e.g., DAITSS[1], JHOVE[2], DROID[3], and the NLNZ metadata extraction tool[4]. Since the development of these applications is based on the metadata schema, we hereby use the OAIS Metadata, the NLNZ Metadata and the PREMIS as examples to illustrate

[1] http://daitss.fcla.edu

[2] http://hul.harvard.edu/jhove/distribution.html

[3] http://www.nationalarchives.gov.uk/aboutapps/pronom/

[4] http://meta-extractor.sourceforge.net/

how well a metadata schema is in terms of our quality requirements. Table 2 gives a comparison evaluation between OAIS, NLNZ and PREMIS Metadata.

Table 2. The Evaluation of the OAIS Metadata, the NLNZ Metadata, the PREMIS. ("+" = "complete", "-" = "incomplete", "n/a" = "no element at all")

Req.	**Items in the requirement**	**The OAIS Metadata**	**The NLNZ Metadata**	**The PREMIS**
R1	• Storage medium profile • Storage driver profile • Usage document	+	n/a	-
R2	• Hardware components profile • Usage document	+	-	-**
R3	• Transfer application profile • Installation document • Usage document	+	-	+
R4	• Interpretation application profile • Installation document • Usage document	+	-	+
R5	• Format technique profile • Specification	n/a	-	-*
R6	• Reference technique profile • Specification	+	n/a	n/a
R7	• Security technique profile • Specification	-	n/a	-
R8	• Content characteristics • Rendering characteristics • Structure characteristics • Behavior characteristics	-**	-	-**
R9	• Changed places • Previous activities	-	-	+
R10	• IPRs • Laws	-	n/a	+
R11	• Preservation level • Process characteristics • Transformation table • Importance factors • Assessment method	n/a	n/a	-

* Having a link to a registry. However, the information in the registry may not be complete. (For example, the PRONOM has no format specification. The MIME Media Type has some standard format specifications, not all format specifications)

** The metadata schema does not explicitly define the necessary elements, but uses a general and repeatable element. Therefore, there exist risks to miss some elements.

The OAIS Metadata is good at providing enough technical background information besides losing specification for the format technique and the security technique's specification. The weaknesses of the OAIS Metadata appear in *R9*, *R10*, and *R11*.

The NLNZ Metadata is incomplete in a word. For hardware, software, and techniques, the NLNZ Metadata just stores names. It implies that the NLNZ Metadata can be used in a short-term preservation system where the related hardware, software, and techniques can be easily found in practice. It also has no elements to describe previous activities and any policy, so that authenticity and provenance are hard to approve. The good place of the NLNZ Metadata is in *R8* where it enumerates content characteristics, but other characteristics are not provided.

The PREMIS is the best schema among these three schemas. Even though *R2*, *R5*, and *R8* are set to incomplete in Table 2, it still has a possibility to become complete as if the preservation system has experienced specialists. The weaknesses of the PREMIS are: 1) no element is used to describe the reference technique, 2) the security technique name is stored but no documentation, and 3) the preservation level can be defined, but the assessment method cannot.

From the case study and Table 2 we may deduce six improvement suggestions for preservation systems:

- It is not enough that metadata schemas hold the names of the hardware, the application, and the specification. Our workflow example showed that more information is required. For example, the usage documentation is needed to reconstruct the hardware environment or operate applications. The specification is needed to develop a new interpretation application or a new conversion application running in a new hardware environment. We suggest therefore that a local registry should be established, which contains complete information about any techniques that are used in the preservation system.
- Some metadata schemas, like the PREMIS, provide links to an external registry. The reliability and completeness of the external registry are risks for the preservation system. For instance, the format registry, Fred (A Format Registry Demonstration), has been stopped, whilst the PRONOM [19] and the GDFR [20] will be combined into one new registry named UDRF [21]. Moreover, the PRONOM currently does not have any format specification, whereas format specifications in the MIME Media Type [22] are messy and incomplete. Therefore, we suggest that an authority organization, e.g., a national library or a national archive, should have a responsibility to register and share technical information, such as hardware documentations, software documentations, format specifications, reference specifications, and security algorithm specifications.
- Techniques may be used to facilitate certain user operations, but on the other hand they increase the risk to the preservation system. For example, the NLNZ Metadata and the PREMIS use the reference technique, but they do not store the related specification. A new application may not parse the references. Therefore, we suggest that as long as a technique is deployed at a preservation system, both the profile and the specification should be preserved.
- Deciding characteristics and setting important degree to the characteristics are two time-consuming and costly tasks. In those two tasks, specialists are invited and sit together to discuss the characteristics and the important degree. Therefore, we suggest that the characteristic and its related important degree should be preserved in the form of metadata. Moreover, it is advantageous for small preservation systems that have few specialists if such metadata are published and defined as a standard.

- A clear retention policy should specify an assessment method to a migration procedure. For example, in the format migration workflow, the numeration rules, the important factors, and the comparison method should be kept for the format migration. In this way, the migration results would always be comparable, and the consistency of the digital objects could be kept. It would also be a basic prerequisite when developing automated migration procedures.
- A tool that automatically generates the migration metadata is important for any migration as the volume of digital objects is too huge for manual operation. Such a tool should have at least two functions. 1) Conversion ability: some metadata (e.g., *R10*-*R11*) obtained from its producer should be converted to the schema supported by the preservation system; 2) Extraction ability: the tool should extract metadata not only about the object's content (e.g., *R8*-*R9*) but also about its technical environment (*R1*-*R7*).

6 Conclusion and future work

In this paper, we pointed out the importance of the quality requirements of migration metadata and assessed whether and to what degree the OAIS Metadata, the NLNZ Metadata and the PREMIS cover the requirements. Those requirements include the hardware, the application, the specification, the object, and the policy. The case study of migration planning indicates that migration metadata of the requirements can play an important role in future migration procedures especially in its optimization.

Our future work is to design 1) a data dictionary to the migration metadata, 2) a user interface that can help curators design a migration metadata schema in terms of their situation, and 3) an extraction tool that automatically extracts and converts the migration metadata.

Acknowledgements. Research on this subject is funded by the Norwegian Research Council and our industry partners under the LongRec[5] project.

References

1. The Digital Preservation TestBed, *Migration: Context and Current Status*. 2001.
2. Lee, K.-H., et al., *The State of the Art and Practice in Digital Preservation.* Journal of Research of the National Institute of Standards and Technology, 2002. **107**(1): p. 93-106.
3. Waters, D. and J. Garrett, *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*. 1996.
4. The Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*. 2002.
5. Wheatley, P., *Migration–a CAMiLEON discussion paper.* Ariadne, 2001. **29**(2).

[5] www.longrec.com

6. The DCMI Usage Board, *Dublin Core Metadata Element Set*. 2008.
7. Thompson, H.S., et al., *XML Schema Part 1: Structures Second Edition.* 2004.
8. Paul V. Biron, K. Permanente, and A. Malhotra, *XML Schema Part 2: Datatypes Second Edition.* 2004.
9. Adler, S., Berglund, A., Caruso, J., Deach, S., Graham, T., Grosso, P., Gutentag, E., Milowski, A., Parnell, S., Richman, J., and Zilles, S., *Extensible stylesheet language (XSL) version 1.0*. 2001, World Wide Web Consortium (W3C).
10. McDonough, J., *METS: standardized encoding for digital library objects.* International journal on digital libraries, 2006. **6**(2): p. 148-158.
11. Lavoie, B. and R. Gartner, *Technology Watch Report - Preservation Metadata.* DPC Technology Watch Series Report 05-01, 2005.
12. Lavoie, B. and R. Dale, *Preservation Metadata and the OAIS Information Model. A Metadata Framework to Support the Preservation of Digital Objects*. 2002, The OCLC/RLG Working Group on Preservation Metadata.
13. The National Library of New Zealand, *Metadata Standards Framework - Preservation Metadata*. 2003.
14. The PREMIS Editorial Committee, *PREMIS Data Dictionary for Preservation Metadata*. 2008,
15. Conklin, J., *Hypertext: An Introduction and Survey*. IEEE Computer, 1987. **20**(9): p. 17-41.
16. Berners-Lee, T., R. Fielding, and L. Masinter, *RFC 3986: Uniform Resource Identifier (URI): Generic Syntax*. 2005.
17. Stephen Grace, G.K., Lynne Montague, *InSPECT Final Report*. 2009, The InSPECT (Investigating the Significant Properties of Electronic Content Over Time) project.
18. Strodl, S., et al. *How to choose a digital preservation strategy: evaluating a preservation planning procedure*. in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. 2007. Vancouver, BC, Canada: ACM.
19. *PRONOM*, Available from: http://www.nationalarchives.gov.uk/pronom/.
20. *Global Digital Format Register (GDFR)*. Available from: http://hul.harvard.edu/gdfr/.
21. *Unified Digital Format Registry (UDFR)*. Available from: http://www.udfr.org/.
22. *MIME Media Types*. Available from: http://www.iana.org/assignments/media-types/.