Standardisation in the area of digital long-term preservation

Sabine Schrimpf; Deutsche Nationalbibliothek; Frankfurt am Main/Germany

Abstract

On request of nestor, the German network of expertise in digital long-term preservation and long-term availability, the German Institute for Standardisation (DIN) has set up a standards committee on long-term preservation standards in the beginning of 2008. This article will introduce the initial work program of the DIN standards committee.

Introduction

With increasing production and use of digital information, the challenge to ensure long-term accessibility of those digital resources, despite the changing conditions arising from rapid technological development, has increasingly gained importance. This challenge concerns memory institutions like libraries, archives and museums, but also public administration, and industrial enterprises. These bodies are faced with the need to preserve digital objects of different kinds – depending on the assignment of the respective organisation for a period ranging from between 5-30 years to several generations of users.

Although legislation is increasingly requesting industrial, cultural, and governmental organisations to ensure the preservation of their electronic records for the long term, the implementation of long-term preservation systems is still in its infancy. For the development, sales and distribution of long-term preservation systems, the existence of accepted standards is of utmost importance. The absence of standards generates insecurity – for potential users as well as for potential providers of long-term preservation solutions. It assists the spreading of individual, local solutions and isolated applications. Yet, nobody can predict whether proprietary solutions of individual providers will still be maintained in a couple of years, or even if the vendor will still be in the market at that time. Therefore, it is important, to agree early enough on disclosed, standardised components, interfaces, formats etc. to be used in digital preservation contexts. Only technical standards can guarantee interoperability and trustworthiness of digital preservation systems in the long run.

This is why most national and international long-term preservation initiatives cover the topic "standardisation" in one way or the other. In recent years, a number of tools and techniques have been developed by various initiatives. Some of them have become accepted as de facto standards in the meantime. Some of the best-known are OAIS, PREMIS, and METS. The history of those three standards illustrates the relative diversity of the long-term preservation landscape.

OAIS (ISO-Standard 14721:2003) [1] is a reference model for an Open Archival Information System. It defines a common methodology and a common vocabulary for long-term archival systems. The development of the standard was initiated by NASA and ESA and drawn up by CCSDS (the Consultative Committee for Space Data Systems). The responsible ISO Technical Committee is ISO/TC 20/SC 13, "Space data and information transfer systems". PREMIS (PREservation Metadata Implementation Strategies) comprises a data dictionary and supporting XML schemas for core preservation metadata needed to support the long-term preservation of digital materials. The PREMIS activities were initiated by OCLC and RLG, while the PREMIS maintainance activities are being hosted by the American Library of Congress. METS, a Metadata Encoding and Transmission Standard, has being developed as an initiative of the Digital Library Federation. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress.

Even for experts, it is difficult to obtain an overview of all the long-term preservation related standardisation bodies and standardisation activities of individual groups. And still, more work on long-term preservation related standards is going on and even greater efforts are needed.

nestor-activities

In this context, nestor, the German network of expertise in long-term preservation and long-term availability of digital resources [2], has a focus on the development and implementation of standards. nestor is a cooperative project of libraries, archives and museums, forming a network of expertise in long-term preservation and long-term availability of digital resources. A goal of the nestor project (funded by the BMBF, German Ministery of Education and Research) is the constitution of a permanent form of organization for all issues of long-term preservation as well as the development of national and international agreements and the assignment of tasks.

Of five nestor working groups, two deal with standardisation: The working group on Digital Repository Certification concentrates on audit and certification issues, and the working group on Long-term Preservation Standards bundles ongoing standardisation activities and advances the establishment of existing standards.

nestor's standardisation efforts are being supported by the German Institute for Standardisation (DIN – Deutsches Institut für Normung). As part of the "Innovation with Norms and Standards (INS) Initiative", supported by DIN and the German Federal Ministry of Economic Affairs, the nestor experts were charged with clarifying the demand for standardisation in the field of digitisation and digital preservation. Several areas of interest with need for action were identified: Persistent Identifier systems, Trusted Repositories/Audit and Certification, Ingest Processes (ingest of digital material into a repository), and Metadata.

In 2007, the DIN welcomed the constitution of a standards committee on long-term preservation standards. This group was constituted in March 2008. nestor experts introduced the following topics there:

- Trusted Repositories Certification
- Quality Management of Trusted Repositories

- Criteria for trustworthy and interoperable Persistent Identifier systems
- Guidelines for Ingest Processes (ingest of digital material into a repository)

Besides these topics, the committee will also integrate the standardisation of legally secure and revision-safe long-term storage of electronic documents and the standardisation of PDF/A. The follwing paragraphs will provide an overview of the scope of the DIN committee and the individual standardisation efforts it integrates.

Standards committee for long-term preservation within the DIN

The DIN committee on long-term preservation resides within the "Information and Documentation Standards Committee" (NABD) [3]. The newly constituted committee merged with the existing committee Archives and Records Management (NABD 15). Traditionally, the NABD is responsible for standardisation of practices relating to libraries, documentation and information centres, publishing houses, archives etc. The NABD work programme concentrates on the development of national and international standards in the field of data elements, transcription and transliteration, library management and performance evaluation, numbering and code systems, permanence and conservation of documents, and archive and records management. With the increasing dissemination of electronic publications and the enhanced use of electronic records and content management systems, ICT related topics make up an ever larger share of its standardisation work items.

At the same time, it is the German national mirror committee to ISO/TC 46, "Information and Documentation" and its sub-committees and working groups. ISO/TC 46 and thereby NABD are already selectively active in the standardisation of digital preservation related topics such as Records Management [4], Digital Object Identifiers (DOI) [5], digital records preservation [6], and ISO/TC 46 has just resolved to establish an ad hoc group to investigate the potention standardisation of the Digital Repository Audit Method based on Risk Management, DRAMBORA. And it can be expected that more long-term related standardisation activities will find their way into the international standardisation organisations. With the DIN committee on long term preservation, Germany has in a timely manner set up a committee to focus and coordinate the national and international standardisation activities in this increasingly more important area.

Trusted Repositories Certification and Quality Management

The nestor working group on Digital Repository Certification engages in the development of a standardised certification process for trusted repositories since 2004. The expert group is working on identifying relevant features (criteria) to evaluate existing and emerging digital object repositories in order to testify their trustworthiness. It addresses the need of establishing a common understanding of what a trustworthy repository is and is not and of defining a process in which the trustworthiness of any repository can be evaluated and certified. Various environments are being taken into account: the library community, the classical archives world, the museums community and also other data producers like governmental institutions, world data centers, research institutions, publishing houses, etc.

Together with software und certification experts, domain experts from the above mentioned environments have been tackling the question of which standards a certification process for digital long-term repositories could be based upon. The result of their work was the nestor "Catalogue of Criteria for Trusted Digital Repositories" [7], published as a draft for public comment in June 2006. The work was conducted in close dialogue with similar international initiatives.

Based on the nestor Criteria Catalogue and similar documents of RLG/OCLC and NARA (namely "Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist" [8]), the international preservation community has recently put in an effort to produce an ISO standard on which a full audit and certification of digital repositories can be based. As a first step, an open discussion group has been created, which is working on a draft document [9]. The aim will be to take this work into ISO in the same way as the OAIS Reference Model, namely via ISO TC20/SC13.

In parallel to the ISO effort, a national certification process shall be standardised by the German Institute for Standardisation. Also, a complementary survey is envisaged, in which existing standards for quality management shall be scrutinised in regard to their applicability for audit and certification of digital repositories.

Criteria Catalogue for Trusted Persistent Identifier systems

Persistent identifiers (PIs) are a central concern in the long-term accessibility of digital resources. Persistent identifier systems have been developed in order to solve the problem of relative instability of traditional web-URLs. The problem with URLs is that they guarantee access only as long as a digital object doesn't change its physical location. PI systems record not only the location, but also the name of a digital object and provide a resolving mechanism in order to ensure retrieval, even if an object's storage location should change. Thus, unlike ordinary web-URLs, PIs guarantee long-term retrieval and referencing of digital resources.

Work on the trustworthiness and interoperability of Persistent Identifier systems is being conducted by the second nestor working group, the working group on long-term preservation standards, in which members of numerous institutions are represented. The initial point of the nestor-survey is the fact that the persistent identifier systems currently in use, such as Uniform Resource Names (URNs), Digital Object Identifiers (DOIs) and the handle system, are generally mutually incompatible. Each system has its own technical and organisational conditions. Institutions looking to set up or use a PI service (e.g. archives, libraries, publishers) therefore need to choose one of the systems. There is, however, no consensus regarding the criteria which users and institutions can adopt for assessing the interoperability and trustworthiness of these systems.

In 2007, the working group compiled a survey on trustworthiness and interoperability of PI systems, which shall serve as a starting point for further standardisation efforts in the DIN committee on long-term preservation issues. The work of the nestor group focused primarily on three aspects:

1. Developing a system which allows PI providers and users to assess objectively the trustworthiness of different PI systems.
2. Taking the special requirements of different institutions (archives, libraries) and PIs into account.
3. Testing a practical solution for the interoperability of resolver systems (Name-to-Thing, „N2T“ [10]).

The concept of trustworthiness, which has already been successfully used in the area of trusted repositories certification, was used on the PI systems when drawing up the criteria catalogue. Criteria were formulated which can be used to test the trustworthiness of different PI systems. The catalogue makes no claim to be exhaustive, however it offers a suitable basis for further discussion. In the course of 2008 the criteria catalogue for trusted PI systems will be presented to an extended circle of experts for discussion.

In two case studies, light was shed on the specific requirements of libraries and archives with regard to PI systems. The case studies show that there are basically a large number of common requirements. The specific requirements arising from the different characteristics of archive stocks and publications are defined separately.

By means of a test implementation of the N2T-concept at SUB Göttingen, an approach was tested, which aims at providing interoperability and support for the trustworthiness of PI systems. As a meta-resolver, it allows access to documents independent from any specific PI-system. The result was that N2T was found to effectively support interoperability and trustworthiness of PI systems. Analysis of the N2T method shows that, besides the aspects already covered, there are further, hitherto largely unresolved issues requiring additional work.

Ingest Processes

The Ingest-process is a critical process in digital preservation, because it initiates the archival process. During the ingest processes, digital objects are transferred into the long-term archival system. Complications that occur during the ingest process may interfere with the quality of all following preservation actions. Moreover, the process involves producers/suppliers of digital resources and the institution that provides the repository. Each of the stakeholders involved has its own role and responsibilities during the ingest process. By clearly defining the relationships between those two stakeholders and standardising the interaction between them, high quality for the preservation process can be ensured from a very early stage.

With PAIMAS (Producer-Archive Interface Methodology Abstract Standard, ISO 20652:2006) [11], a related standard already exists. It remains at a very high level, though, and defines ingest-interactions on an abstract and rather conceptual level. The merit of PAIMAS is that it provides a methodology and a vocabulary for further work in this area. In the last chapter of the PAIMAS standard, the need is addressed to produce more concrete and application-oriented Producer-Archive Interface Methodology Community Standards from the Abstract Standard.

The nestor standards working group has begun to develop such a community standard for the requirements of memory institutions (archives and libraries). It aims at defining a manual that assists archives and libraries in the endeavour to implement feasible ingest-workflows for their archival systems.

Legally secure and revision-safe long-term storage of electronic documents

Public administrations and industrial enterprises increasingly rely on electronic documents. The legislation has recognised this fact and defined a regulatory framework, which place electronic records on a par with conventional paper records.

When it comes to the legally secure and revision-safe long-term storage of electronic documents, there is still some insecurity. Because digital records can easily be manipulated, it is essential to implement measures that ensure the authenticity and integrity of the stored documents. According to the state of technology, electronic signatures guarantee authenticity, integrity, confidentiality and completeness of electronic documents.

Actually, some international, European and national specifications for electronic Signatures exist (e.g. CMS, XML-DSig). Though, those specifications do not refer explicitly to the long-term storage of electronically signed documents. But exactly the long-term storage of electronically signed documents causes some difficulties, due to the following three reasons:

1. Since electronic signatures become less secure with the time, electronic records have to be re-signed periodically.
2. Verification data has to be stored with the document for the long-term, too.
3. When formats of electronic documents become obsolete, the signed document must be migrated to a new format in a legally secure and revision-safe way.

Approaches to handle those challenges were developed in the projects ArchiSig [12], ArchiSafe [13] and Transidoc [14]. Concertedly, they were aiming at integrating electonic signature components into long-term archival systems and thus introducing uniform standards for the legally secure and revision-safe long term storage of electronic documents all over Germany. Furthermore, the ArchiSig solution for a legally secure long-term preservation of electronically signed documents resulted in an IETF (Internet Engineering Task Force) Request for Comments: RFC4998 on „Evidence Record Syntax“ [15]. The basis of the ERS are Archive Timestamps, which can cover a single data object or can cover a group of data objects by using hash trees, combined with a timestamp. The merit of this concept is that the deletion of a data object in the tree does not influence the verifiability of others.

The DIN committee constitutes an adequate committee to formally recognise national standards and to contribute to further international standardisation activities in the context of legally secure, revision-safe long-term preservation.

PDF/A

The PDF/A standard ISO 19005-1:2005 [16] is based on the PDF Reference Version 1.4 from Adobe Systems Inc. Adobe handed its PDF (Portable Document Format) specification over to ISO/TC 171, Document Management Applications in order to allow the definition of a PDF specifiation suitable for archiving and preserving documents. It thereby addresses one of the major challenges of long-term preservation: the threatening obsolescence of a multitude of currently existing formats. Rapid technology change will most probably prevent documents saved in certain formats from being reproduced in the exact same way some time in the future.

With the ISO standard PDF/A, a standardised, stable, open format specification has been created, “a file format based on PDF,

known as PDF/A, which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independently of the tools and systems used for creating, storing or rendering the files." [17] In order to reach this goal, PDF/A leaves out some features of PDF not suitable for long term preservation. For example, the integration of audio and video content, JavaScript and executable file launches are forbidden. Instead, the integration ("embedding") of all fonts and other information necessary for displaying the document is prescribed. This includes all content (text, raster images and vector graphics), fonts, and color information. That way, it is ensured that PDF/A documents can be reproduced in exactly same way as they were intended to be.

The specification will be advanced and administered by ISO/TC 171. Adobe announced in January 2007 to hand over PDF 1.7 for standardisation soon. The standardisation experts' task will be to ensure the compatibility of the new version PDF/A-2 with the first one in order to avoid what shall ultimately be solved with PDF/A: unnecessary migrations and obsolete formats.

Outlook

The discussion of the topics in this paper is still far from conclusive. As mentioned above, standardisation in the area of long-term preservation is in its early stages and more topics are likely to come up in the near future.

References

[1] ISO 14721:2003 Reference Model for an Open Archival Information System (OAIS)

[2] www.digitalpreservation.de

[3] http://www.nabd.din.de/cmd?search_committee=nabd&workflowname=InitCommittee&contextid=nabd&languageid=en

[4] ISO 15489-1 Information and documentation – Records management – Part 1: General, and ISO/TR 15489-2 Information and documentation – Records management – Part 2: Guidelines

[5] ISO/CD 26324 Information and documentation – Digital Object Identifier (DOI)

[6] ISO/DTR 26102 Information and documentation – Requirements for long term preservation of electronic records

[7] Catalogue of Criteria for Trusted Digital Repositories – Version 1 (draft for public comment), December 2006. URL: http://edoc.hu-berlin.de/series/nestor-materialien/8/PDF/8.pdf

[8] Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, Version 1.0, February 2007. URL: http://www.crl.edu/PDF/trac.pdf

[9] http://wiki.digitalrepositoryauditandcertification.org/bin/view

[10] http://n2t.info/

[11] ISO 2005:20652 Producer-Archive Interface Methodology Abstract Standard.

[12] http://www.archisig.de/

[13] http://www.archisafe.de/

[14] http://www.sit.fraunhofer.de/projekteundthemen/transidoc.jsp

[15] RFC 4998 Evidence Record Syntax (ERS). URL: http://www.ietf.org/rfc/rfc4998.txt

[16] ISO 19005-1:2005 Document management – Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1)

[17] ISO 19005-1:2005

Author Biography

Sabine Schrimpf is a scientific assistant in nestor at the DNB. She received her master's degree in publishing sciences at the University of Mainz. She studied Publishing sciences, American studies and Literature at the University of Mainz. At present, she actively participates in nestor, the network of expertise in long-term storage of digital resources for Germany