LIBER QUARTERLY
THE JOURNAL OF THE ASSOCIATION OF EUROPEAN RESEARCH LIBRARIES

Vol. 23, no. 3 (2014) 201–213 | ISSN: 1435-5205 | e-ISSN: 2213-056X

nestor Guideline for Preservation Planning – a Process Model

Sabine Schrimpf

Deutsche Nationalbibliothek
s.schrimpf@dnb.de

Christian Keitel

Landesarchiv Baden-Württemberg
Christian.Keitel@La-bw.de

Abstract

The nestor guideline for preservation planning is the latest in a series of nestor publications. nestor is the German competence network for digital preservation and it offers all interested parties from the private and public domains the possibility to participate in working groups. The guideline for preservation planning is the result of such a working group, which discussed the conceptual and practical issues of implementing the OAIS Functional Entity "Preservation Planning".

The guideline describes a process model and offers some guidance on potential implementations. It integrates and builds on recognized community concepts like Significant Properties, the OAIS Designated Community, the National Archives of Australia's Performance Model, the PREMIS concept of Intellectual Entities and Representations, and the PLANET's approach to preservation planning. Furthermore, it introduces the concepts "intended use" (Nutzungsziele), "information type" (Informationstyp) and "preservation group" (Erhaltungsgruppe). The purpose of these new categories is that information objects shall be grouped by information type (e.g.,

Igitur publishing | http://liber.library.uu.nl/ | URN:NBN:NL:UI:10-1-116062

audio, video, text…) and intended use (e.g., reading for pleasure, search for specific information…) to preservation groups for automatic processing. Significant properties can then be derived for whole preservation groups. The file format alone is considered as not completely sufficient for such categorisation. Some exemplary implementation solutions of the new concepts are presented in an annex.

The guideline takes into account that resources for preservation planning and preservation actions are limited and has therefore adopted 4 premises: adequacy, financial viability, automation, and authenticity of archived objects. Its pragmatic approach becomes apparent in the definition and explanation of these dimensions. The guideline is written from the point of view of representatives of memory institutions, i.e., libraries, archives, museums, and is primarily targeted at this context although it may be useful for other information preserving institutions too.

This contribution introduces the nestor guideline for preservation planning (for now only available in German; English translation envisaged for 1st half of 2014) to an international audience for the first time. It also matches the process model and the new concepts intended use, information type and preservation group to the collection and preservation reality of the German National Library.

Key Words: digital preservation; preservation planning; significant properties; LTP-policies; workflow optimisation

1. Introduction: The nestor network and its working groups

nestor, the German network of expertise in digital preservation, was set up in 2003 in recognition of the fact that digital preservation is a task too big to be solved by any single institution. The mission of the network is to bring together experts in digital preservation, to foster knowledge exchange and networking, and to provide information, expertise, training and a forum for standardisation opportunities to the interested communities. Since its beginning, nestor has offered working groups on different preservation-related topics to all interested parties from the private and public domains. One of the latest Working Groups (WG), the WG on Digital Preservation, set up in 2009, dealt with the question of how to plan preservation measures for intangible digital assets. In contrast to books or paper-based archive records, digital

objects can only be preserved by means of proactive measures on the part of the archive or library. Due to a lack of longstanding experience in the relatively new field of digital preservation, however, it is hard to tell which measures are appropriate in the first place. The working group brought together experts from the archives, libraries, and museums sector. They discussed, for example, at what point in the preservation lifecycle certain assumptions can or must be made and certain actions can be initiated. They investigated how user interests, significant characteristics and practical requirements can be reconciled. Finally, they published the nestor guideline for preservation planning as a result of their work (Nestor Arbeitsgruppe, 2012).

2. The nestor guideline for preservation planning

2.1. Relevant work

The working group started with extensive desk research and discussion of recognised community concepts:

The term "Significant Properties" was first used by the CEDARS project in 1999 (Cedars Project, 2002) and the concept was then widely discussed within the preservation community. It refers to those characteristics of a digital object that must be maintained over changes in technology and potential file format migrations.

The Reference Model for an Open Archival Information System (OAIS), first published in 2002 and revised in 2009 (CCSDS, 2009), introduced several key concepts. The OAIS Functional Entity "Preservation Planning" and the concept of a "Designated Community" were considered most relevant for the working group's questions. The Functional Entity "Preservation Planning" encompasses tasks such as development of preservation strategies and standards, development of packaging designs and migration plans, and monitoring of technology and the Designated Community. The Designated Community is in OAIS defined as an identified group of potential consumers of digital information who should be able to understand a particular set of (preserved) information.

The National Archives of Australia's Performance Model (Heslop, Davis, & Wilson, 2002) is central because it systematically breaks down the concept

of a digital record into fundamental components: The source (the data file), the process by which it is mediated (hardware and software) and its performance (when it is rendered on a screen). The goal of any preservation action is always the preservation of the performance, even if the source has to be changed (i.e., migrated) in order to reach the goal. The determination and preservation of the "essence" (another term for Significant Properties) of any given digital object becomes a key activity in this context.

Finally the PREMIS data model introduced the concepts of Intellectual Entities, Objects and Representations (PREMIS, 2002). Thereby, the Intellectual Entity is the intellectual work that can be described as a whole with properties such as author, title, publication date. It can be manifested in several representations, e.g., as a text and an image file. Each representation can consist of several objects, e.g., one digitised book can consist of hundreds of individual images.

The PLANETS approach to preservation planning (Strodl, Becker, Neumayer, & Rauber, 2007) builds to some extent on the work described above. However, it is not trivial to operatively build it into regular archival business routines.

These concepts have, more or less, existed side-by-side so far so that each digital archive had to take the decision which aspects it needed to integrate how in its own systems and working routines. The nestor guideline on preservation planning now draws all of these concepts together and integrates them in a single process model. Thereby, it intends to describe a pragmatic approach that is easy to implement in small scale as well as in large scale institutions.

2.2. Key assumptions and concepts

With unlimited resources, the digital preservation challenges were more easily manageable. The available resources, however, restrict the scope of action and often require a prioritisation of preservation goals. When dozens of original production formats cannot be supported, the archive must make a decision about suitable preservation formats. When that means that not all archival holdings can be preserved with their full original functionality, the archive must take decisions about the most preservation worthy features and functionalities. Each decision for something is also a decision against

something else. In order to acknowledge these framework conditions, the nestor guideline has adopted 4 premises:

1. Financial viability: Digital preservation and the related preservation planning must be economically affordable.
2. Adequacy: Preservation goals must be adequate for the particular archival institution and may, for the same type of digital information, differ between institutions according to their preservation mandate or their designated community.
3. Authenticity: The goal of any preservation action must always be to maintain the authenticity of all archived objects. If (future) users cannot trust the archival holdings, all preservation efforts are in vain.
4. Automation: The sheer amount of digital objects requires that archival objects are processed group-wise and as automatized as possible.

To support the decision making, which is inseparably linked with proactive preservation planning in times of restricted budgets, the guideline introduces three new concepts:

- "Intended use" is used to describe for what purpose the designated community will use the archived information, or with which questions it will approach them, e.g., reading for pleasure, search for specific information, …
- "Information type" describes the type of information, e.g., audio, video, text, …
- "Preservation groups" are created when information objects are grouped by information type and intended use.

The idea behind the new concepts is that information objects shall be grouped by information type and intended use to preservation groups for subsequent automatic processing. Significant properties can then be derived for whole preservation groups. The file format alone is considered insufficient for such categorisation because it hardly relates to the intellectual content of the information object. For example, a collection of avant-garde digital photographs with multiple different imaging formats may have more commonalities in terms of intended use and preservation goals than a stock of various digitized text and image records that happen to be saved in the same digitisation format.

2.3. Process model – Initial information ingest

In accordance with the Performance Model, the goal of any preservation action is the preservation of the performance of any given digital record. Ideal-typically, a digital archivist perceives the performance as it is rendered from the data source when a record is submitted from the producer to the archive. At that moment, a certain combination of hardware and software is available in order to recreate the original performance. The digital archivist cannot "see" the data source or the software operations needed for the rendering process, he can only perceive the resulting performance. He can, for example, look at an image or text that is rendered on a screen or listen to sound from speakers or earphones. Thereby, the archivist perceives all features of the rendered performance, and he can draw conclusions as to the underlying information objects. Some features might be more obvious that others, e.g., the order of words in a text or the colour of a graphic, some might be less tangible, e.g., the font, the underlying colour space of an image or the tonal space of a music performance.

However, as the original hardware and software that creates the performance is likely to change over the years, and as in addition the data source might be converted to new file formats, it must be assumed that not all original features of the archived record can be preserved forever. It is the task of the digital archivist to determine the essential characteristics, the so-called significant properties, which will become the benchmark for future preservation actions. As the significant properties are a means to enable meaningful future use of the archived holdings, the archivist ought to start from the requirements of the designated community and its intended usage scenarios (who will use the archived information in the future, with which preconditions, for what purpose?).

In order to deal with large amounts of digital information objects, the archivist can group information objects with similar features and the same designated communities and intended uses together into preservation groups to determine their significant properties. The affiliation of information objects with preservation groups is recorded in the metadata of the information object.

2.4. Process model – Creation of preservation groups

Ideal-typically, the creation of preservation groups starts with the definition of the information types that the archive has to deal with (e.g., audio, video,

etc.), followed by the definition of designated communities and intended uses per information type. Thereby, it is up to the archive how detailed it wants to describe and characterize its designated communities. It could, for example, leave it with "historian", but also go into more detail, e.g., economic historian, medievalist etc. The characterisation of the designated community could, among others, take into account aspects like the level of expertise about content and technology, standard technical equipment, legal restrictions, or the size of the designated community. The same holds true for the description and characterisation of the intended uses. The nestor guideline proposes four main cases as starting point:

1. Perception of the work
2. Analysis of the work/information retrieval
3. Further processing of the content
4. Execution of the item/running its applications (just for software)

Thus, a variety of qualified subsets of the rather broad information type groups are created. They are the preservation groups. Based on exemplary performances rendered from information objects of a specific preservation group, the archivist gets an idea of the characteristics of the underlying information objects within the group. It is clear that there is a dependency and reciprocal effect with the definition of intended uses: The archivist must, possibly by previous rendering, have ascertained that the intended use corresponds with the possible use of an information object. A still image, for example, cannot be executed in the same way as a computer game.

Depending on the intended use of the objects by the designated community, the archivist derives their significant properties from all features that characterize the objects within the group. Finally, the degree of necessary fulfilment of the significant properties is determined per group. The significant properties and their degrees of performance are not static but must be adapted as the designated communities and also the intended uses change over time.

The preservation groups as such do not need to remain static over time. As the requirements concerning information objects change, preservation groups may have to be restructured, split or merged. The preservation groups have the benefit that they lay, early in the preservation lifecycle, the foundations

Table 1: Example of a preservation group matrix for text materials: The top row lists various characteristics and the left column the intended uses. Next to them, the significant characteristics are marked and the right column assigns an appropriate file format. The information in columns 1-5 can be used in each case a new file format is needed.

	Text font	Order of words	Full-text search	Mark-up	Suitable format today	Potential future format
1) Text reading		x			TIFF	... (tbd)
2) Text analysis		x	X		PDF/A	... (tbd)
3) Further processing		x	X	x	XML	... (tbd)

for a smart, since differentiated, automated group-wise processing of the archival holdings. The obvious grouping by file format alone seems rather undifferentiated in comparison and carries the risk that significant properties can only represent the lowest common denominator.

3. Preservation Planning

Again, the idea of the preservation of the performance of a digital record guides all considerations concerning preservation planning. The underlying data stream and/or the hardware and software that renders it will sooner or later be subject to change. For preservation planning purposes, the archive ought to take several actions.

3.1. Monitor Designated Community and Technology

Most importantly, the digital archive must make sure it recognizes when its archival holdings are threatened by obsolescence. For that purpose, it monitors the general technological development, and of equal importance, the designated community as it is the most important reference for the archive. When the technical prerequisites of the designated community change and it becomes apparent that this will affect their use of the archive's information objects, the archive must react to it and bridge the gap for the designated community. It can do so by accumulating and providing sufficient representation information to the designated communities or by initiating preservation actions.

3.2. Preservation Strategies: Migration and Emulation

Of all preservation strategies, migration and emulation are best understood and widely accepted. The concepts in the nestor guideline can be integrated with both concepts. Moreover, they help to conceptually compare the capacity of the emulation and migration strategy for any given preservation group.

Preservation actions are planned and executed preservation group-wise. When the archive decides to migrate information objects to new file formats, it must identify a target file format that supports the intended uses comparable to the original format. The recorded significant properties act as a benchmark to evaluate the success of the migration action and, as a migration always changes the underlying information object, to retain the authenticity of the migrated objects. The values of the significant properties of the new information object and its performance are recorded again and compared with the values of the significant properties of the original object. The results of the comparison are documented and stored with the preservation group's objects.

In order to perform the emulation strategy, an emulator must be acquired or developed that supports the intended uses of a preservation group. Similar to the procedure described for migration, the recorded significant properties act as benchmark. In order to evaluate the success of the emulation action, the values of the significant properties of the emulated information object and its performance are recorded and compared with the values of the original object. The results of the comparison is documented and stored for the long term.

4. Mapping the nestor guideline with the long term preservation routines in the German National Library

The starting point for the digital preservation activities of the German National Library (DNB) was the kopal project (2004–2007), in which a long term archival system based on IBM DIAS was developed. For data management, it makes use of the specifically developed Universal Object Format (Steinke, 2006), which allows for archiving of digital objects along with preservation metadata. The preservation system was enhanced in the DP4lib

project (2009–2012) and the long term preservation workflows, which were previously rather isolated, were integrated with the established online publication collection routines. In the first quarter of 2013, the nestor guideline for preservation planning were used as a benchmark to evaluate the current status and to reinforce the basic preservation planning approach.

4.1. Information ingest at the DNB

Digital publications are processed entirely automatically at the DNB. Some descriptive metadata for each information object is supplied by the publishing companies and other information providers. Technical metadata is generated automatically on file level (one information object can consist of multiple files).

Each information object is automatically assigned to a so-called "object group", which is roughly, although not exactly, comparable to the nestor guideline's *information type*. Examples of object groups include audio book, e-book, e-paper, online dissertation, print-on-demand, journal, journal article, digitisation, website. It is conceivable that these object groups could be refined according to intended use and user group, e.g., "all e-books with multimedia content", or "all scientific audio books". Thus, preservation groups could be created as subsets of the DNB object groups.

Significant properties are currently not recorded during the ingest process. They could, however, be recorded on the object level as well as on the file level as the metadata formats allow that.

4.2. "Preservation groups" at the DNB

The data management structures in the long term archive do not allow group-wise organisation of archival holdings according to intellectual criteria like the affiliation to a preservation group, because it does not hold descriptive metadata. The archival database takes a mere technical view on archival objects and allows selection according to technical characteristics that are recorded as part of the technical metadata. This could be, for example, "all PDF versions older than PDF 1.4", or "all TIFF files within the object group digitisation". In the event of a migration project, files are selected according to these technical criteria.

The selection and group-wise treatment of preservation groups involves a small detour: Via the library catalogue which holds all descriptive metadata, all objects of a specific object group (or, in perspective, preservation groups as potential subsets of an object group) can be selected. In this case, a list of the URNs of the related information objects may be created and on the basis of the URNs, the objects may finally be retrieved from the long term archive for group-wise treatment.

4.3. Preservation Planning at the DNB

The DNB takes a threefold approach to the monitoring tasks: It has made provisions to monitor the technical fitness of the long term archive itself, to do regular risk assessments of the stored object types (which includes technology monitoring for the said object types), and to plan for preservation actions. All three strands can and will in the future be conceptually supported by the nestor guideline, especially in reference to monitoring preservation groups, conceptually or technically.

Migration and emulation exist, so far, only as conceptual strategies, or rather, migration has been executed as a proof of concept. Because no significant properties are recorded at the time of ingest, the designed migration process differs from the one outlined in the nestor guideline. Again, the DNB takes an event-based approach: The significant properties of the selected information objects are not determined until the decision for migration is taken. Depending on the significant properties, a target file format is selected. A migration tool is selected and tested paying attention to the preservation of the significant properties. After successful tests, the migration is executed automatically and only the significant properties of samples are compared.

4.4. Conclusion of the mapping exercise

Although the DNB could not completely map to the nestor guideline, the concepts and outlined process models were found very useful as a benchmark. As a result of the mapping a gap was revealed with regard to significant properties because they are, conceptually, only determined in the case of a migration event. This has two disadvantages: (1) they cannot support preservation planning, although their preservation should be the goal of any preservation action; (2) it may be too late to save some of the significant

characteristics when they are determined retrospectively. The creation of preservation groups, or the determination of significant properties, at least on object group level, is therefore perceived as an ideal approach to determining significant properties relatively pragmatically without deeply altering existing data management structures.

5. Outlook

The nestor guideline were publicly presented and discussed several times at national level, most recently at a workshop at the German Librarians' Day in March 2013. The audience valued the concepts and considerations of the guideline and unanimously considered it a good starting point for implementing preservation planning at the institutional level. Several participants expressed the wish, however, for more practical guidance and experience sharing. In addition to theoretical examples of preservation groups they would find a knowledge base of tried and tested preservation groups with file format recommendations valuable. This goes slightly beyond the scope of the published guideline but is certainly an interesting perspective to follow up in nestor if or once a critical mass of interested parties has started to implement the concepts of the guideline.

References

CCSDS (2009). *Reference model for an open archival information system (OAIS). Draft recommended standard CCSDS 650.0-P-1.1.* The Consultative Committee for Space Data Systems. Retrieved January 14, 2014, from http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf.

Cedars Project (2002). *Cedars guide to digital collection management*. The Cedars Project. Retrieved January 14, 2014, from http://www.webarchive.org.uk/wayback/archive/20050111120000/http://www.leeds.ac.uk/cedars/guideto/collmanagement/guidetocolman.pdf.

Heslop, H., Davis, S., & Wilson, A. (2002). *An approach to the preservation of digital records.* Canberra: National Archives of Australia. Retrieved January 14, 2014, from http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf.

Nestor Arbeitsgruppe (2012). *Leitfaden zur digitalen Bestandserhaltung*, Version 2.0. Frankfurt am Main : nestor c/o Deutsche Nationalbibliothek. Retrieved January 14, 2014, from http://nbn-resolving.de/urn:nbn:de:0008-2012092400.

PREMIS Editorial Committee (2002). PREMIS data dictionary for preservation metadata, version 2.0. Retrieved January 14, 2014, from http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.

Steinke, T. (2006). Universal Object Format. An archiving and exchange format for digital objects. Frankfurt am Main: Project kopal. Retrieved January 14, 2014, from http://kopal.langzeitarchivierung.de/downloads/kopal_Universal_Object_Format.pdf.

Strodl, S., Becker, C., Neumayer, R., & Rauber, A. (2007). How to choose a digital preservation strategy: Evaluating a preservation planning procedure. JCDL 2007 Proceedings (pp 29–38). New York, NY: ACM. Retrieved January 14, 2014, from http://www.ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf.