iPRES 2009

THE SIXTH INTERNATIONAL CONFERENCE ON THE PRESERVATION OF DIGITAL OBJECTS

Proceedings

October 5-6, 2009
Mission Bay Conference Center
San Francisco, California

California Digital Library

Into the Archive: Potential and Limits of Standardizing the Ingest

Jens Ludwig
Niedersächsische Staats- und Universitätsbibliothek, Papendiek 14, 37073 Göttingen
Jens Ludwig, ludwig@sub.uni-goettingen.de

Abstract

The ingest and its preparation are crucial steps and of strategical importance for digital preservation. If we want to move digital preservation into the mainstream we have to make them as easy as possible. The aim of the NESTOR guide "Into The Archive" is to help streamlining the planning and execution of ingest projects. The main challenge for such a guide is to provide help for a broad audience with heterogeneous use cases and without detailed background knowledge on the producer side. This paper will introduce the guide, present first experiences and discuss the challenges.

The importance and complexity of ingest

For a number of reasons the ingest and its preparation are of special importance for digital preservation.

As Beagrie, Chruszcz and Lavoie (2008) have documented, ingest can be regarded as cost factor number one for digital preservation. For example, for the social sciences oriented UK Data Archive ingest is responsible for 42 percent of all costs while access makes up 35 percent and archival storage and preservation even only 23 percent. This is cost-intensive, but it prevents more expenses at a later stage: The Digitale Bewaring Project (2005) in the Netherlands estimated that metadata creation ten years after the ingest costs about 30 times more than doing it right at the beginning.

The ingest is also the first step for digital preservation and most often the first question of people interested in digital preservation: How can I transfer my object to you? This is of course a legitimate question but often the questioner is not aware that a simple copy process is not sufficient. There is a major need for information on the side of the producer.

Most importantly ingest has to ensure a sufficient quality and the possibility of future reuse of the digital objects. The ingest decisions and procedures can have serious consequences for all future activities and errors in this process may be irreparable. This is deeply connected to the complexity of ingest.

Unfortunately the ingest process is not only important, but also quite complex. The complexity of the ingest process is an inherent attribute of digital preservation. Data needs to be transferred from heterogeneous and organisation-specific contexts in such a way that it will nevertheless remain comprehensible and reusable in different contexts in the future. To archive this reusability the implicit usage context needs to be made explicit and the preservation requirements need to be defined. But to know exactly what context information needs to be captured and which preservation requirements are necessary would require to foresee the future usage.

This is also the main point why the ingest into a long-term archive should be distinguished from the ingest into a normal repository. The repository community also sees an ingest challenge and tries "Breaking the Repository Ingest Barrier" (Yeadon 2008). But the various efforts in that area aim primarily to simplify the deposit from users in order to motivate more deposits. The question of preservation requirements is usually not addressed (and this is not necessary).

Standardization and guidelines

The complexity of the ingest is a major obstacle for simplifying this strategically important process. A standardization of the long-term preservation ingest would help streamlining the process but this can not simply be a general prescription of technical procedures and interfaces. The use-cases for ingest and the underlying technology are too diverse and changing too fast. Additionally the main time-intensive tasks connected to the ingest are not technical, but communicational and organisational.

Since the ingest is important and complex and information about it is often requested, the NESTOR working group for long-term preservation standards decided to design a guide which clarifies the goals and unique aspects of ingesting information into a digital archive on a high level. It tries to provide an introduction and common working basis to memory institutions and information producers/providers for planning the ingest so that their cooperation can proceed as smooth as possible.

The NESTOR working group consisted primarily of persons active in libraries and archives, but the guideline is not only intended for memory institutions. Apart from the funding by the German Federal Ministry of Education and Research the work was also supported by the Initiative "Innovation with Norms and Standards" of the Federal Ministry for Economics and Technology and the DIN, the German Institute for Standardization.

The development of the NESTOR Guide "Into The Archive"

In order to produce a guide as useful as possible, the NESTOR working group for long-term preservation standards made a number of decisions during the development:

1. Not to use the full OAIS terminology (but a mostly compatible one)
2. To define ingest as transfer of responsibility
3. To organize the process in small and manageable sections

The decision not to use the full OAIS terminology is simply based on the experience that it is quite difficult for people unfamiliar with the OAIS. Using its terminology would make it more difficult to keep an introductory character. But since the merits of the OAIS reference model are undeniable, special attention was paid to be compatible to it.

Especially worth mentioning is that the definition of ingest was not adopted from the OAIS. The OAIS defines "ingest" as a functional entity of a digital preservation system and not as an outcome. It is finished when all processes to store the AIP are finished. Instead, the NESTOR guide defines "ingest" as "the organisation and execution of all processes necessary to accept an information object into the archive and for the archive to assume responsibility for it" (nestor 2009). This excludes some activities of the OAIS, since the guide tries to leave out those activities which are only relevant for the internal operations of the long-term archive. Moreover it includes additional activities which are not part of the OAIS like appraisal or the definition of authenticity requirements.

Of course the Consultative Committee for Space Data Systems has already published the Producer-Archive Interface Methodology Abstract Standard (PAIMAS) which deals with the preparation of the ingest by producers and archives. This is a very helpful standard and was very important for the development of the nestor guide. But in the opinion of the working group it is with it 87 very granular steps too complex, daunting and not compact enough for non-experts. For these kind of tasks less granular guidelines may be more helpful since they make it easier to see the big picture.

It may surprise that the ingest tasks in the guide are presented in an order in which they have to be addressed in practice. The reason for this is that in practice only few assumptions can be made about the order in which these tasks have to be dealt with. The complex real-life dependencies require that tasks have to be dealt from case to case in different order, sometimes simultaneously, sometimes repeatedly. For example, tasks like identifying the optimal balance between the amount of data to archive, the therefore possible and necessary transfer methods and the resulting costs will often require an iterative procedure. Instead of prescribing an order of operations the guide groups tasks in three thematic blocks with three topics. The tasks are explained not as a set of instructions but as objectives with example procedures.

Structure and Content of the NESTOR Guide "Into The Archive"

All in all this lead to the following simple structure of ingest targets:

1. Objects
    a. Information to be archived
    b. Metadata
    c. Significant properties
2. Processes
    a. Transfer packages
    b. Validation
    c. Transfer of data
3. Management
    a. Laws and contracts
    b. Ingest agreement and documentation
    c. Costs and staff

The following paragraphs briefly present the way the topics are explained in the guide and some additional aspects worth mentioning.

Objects

This is the subject matter producers have usually the best conception of, but still it has to be clearly communicated what the archived objects really are. There are three essential tasks relevant to the ingest of digital objects: The information to be archived has to be selected and the necessary metadata and the significant properties have to be defined.

The selection of the information to be archived has to cover the intellectual and technical aspects. Usually the intellectual entities are the objects which are of interest and their possible technical export form is secondary as long as it satisfies the preservation requirements.

The kind of metadata necessary and how much of it is needed depends on the later usage context. Producers often think only about descriptive metadata since it is the most and probably only well-known type of metadata. The other types of metadata producer and archive have to discuss are of course those of the OAIS and PREMIS but esp. semantic and syntactic representation information.

It is necessary to already clarify at the beginning which properties are should be preserved. Fortunately, lately some work has been done in the area of significant properties, namely by the InSPECT-Project or in the context of the PLANETS preservation planning tool PLATO. The documentation of the significant properties allows the definition of a qualitative preservation level.

Processes

The essential tasks regarding the processes related to the ingest are the definition of the transfer packages (the SIP in OAIS terminology), of the validations and of the transfer process itself.

For the producer system and the long-term archive the transfer packages are a kind of common language. These will internally manage information objects in different ways and therefore a "translation" is necessary as an intermediate step.

After the transfer validations have to take place to ensure the correct transfer. The tests for integrity,

technical validity and completeness are usually standard, but also checks for semantic validity could be agreed upon (e.g. if a specific value of a scientific calculation is in a predefined interval). But probably the most important task is not to conceive tests for even the most rare errors but to clarify the required degree of compliance and the consequences if a test is not passed. If no valid version of the object is available or the workflow for non valid objects can not deal with large quantities strict tests can not raise the quality.

For the transfer itself a number of options exist. Here again the main task is often not the definition of the technology, but the definition and testing of the transfer steps, the workflows and feedback mechanisms, legal and security requirements, etc.

Management

The management of the ingest requires dealing with a lot of topics, but the working group decided that an introductory guideline mainly has to include the laws and contracts, ingest agreements and documentation, and costs and staff (Right now the third pair is actually not costs and staff, but a section on all kinds of management activities which breaks with the structure of the other tasks. This is an artifact of the development process and will be changed.)

Legal and contractual questions are of course very relevant for long-term preservation. Intellectual property rights and copyright rights can prevent the necessary modification and replication of the object. But if ingest is defined as transfer of responsibility then legal and contractual documents can also become enabler of preservation. They can create trust beyond good faith if the obligations, preservation levels and liability are defined in them (resp. in appendixes).

The planning decisions as well as the actual process have to be documented. This documentation provides a guideline in implementing the ingest and in resolving issues during the process. But more important, it is also the only instrument to audit the ingest and the authenticity and integrity of the objects. For trustworthiness the transparent and citable documentation is indispensable.

And, of course, the costs have to be addressed. There are already some models to estimate the costs of long-term preservation and especially the costs for all other ingests tasks can be estimated. Many people are discourage by the costs of the ingest, which is as stated above the biggest part of the preservation activities. Often it is sensible to take also the costs of *not* archiving into account, if that is an option at all, to put things into perspective.

Discussion, Feedback and Experience

First discussions with potential users have shown that there is much interest, and especially the briefness was regarded as positive. Suggestions for further developments have been to include the viewpoints of more domains (e.g. museum, industry, science) and to provide more examples, rationales and use cases. A new version of the NESTOR guide Into the Archive will be developed and experts from those domains will be asked to proofread the guide.

Another insight gained in exercises was that one should not neglect the people involved in the ingest preparation. Ideally they already have an understanding of the task and technical and legal expertise. For a good cooperation it is necessary that they have a common goal with the ingest and that long-term preservation is not just seen as unwanted obligation. This will be an additional aspect in the next version.

References

Beagrie, N.; Chruszcz, J.; and Lavoie, B. 2008: Keeping Research Data Safe – A Cost Model and Guidance for UK Universities. JISC 2008. http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Consultative Committee for Space Data Systems (CCSDS) 2004: Producer-Archive Interface Methodology Abstract Standard (PAIMAS). http://public.ccsds.org/publications/archive/651x0b1.pdf

Consultative Committee for Space Data Systems (CCSDS) 2002: Reference Model for an Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0b1.pdf

Nationaal Archief 2005: Costs of Digital Preservation. The Hague. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf

nestor working group for long-term preservation standards 2009: Into the Archive - a guide to the information transfer to a digital repository (Draft for public comment), http://nbn-resolving.de/urn:nbn:de:0008-20080710002 (German Version: nestor AG Standards: Wege ins Archiv. Ein Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv. nestor 2008. URN: urn:nbn:de:0008-2008103009)

Yeadon, S. 2008: Breaking the Repository Ingest Barrier. In: Third International Conference on Open Repositories 2008, 1-4 April 2008, Southampton, United Kingdom. http://pubs.or08.ecs.soton.ac.uk/10/