File Format Migration as Digital Preservation Strategy – A Report on Practical Experiences

Alexander Herschung, Martin Oelgeklaus

startext GmbH, Dottendorfer Straße 86, 53129 Bonn, Germany
Alexander.Herschung@startext.de, Martin.Oelgeklaus@startext.de

Abstract. The Open Archive Information System reference model is a first solution for handling the big number of digital documents. With startext SORI we created a ready-to-use software, based on the OAIS model. So files get prepared for the future, but like every model, there are still aspects to improve.

Keywords: Digital Archive, OAIS, Long-Time Preservation, Software

1 Introduction

In times where paper is a less used resource and more and more documents are created and stay in a digital form, it gets more and more important to get a practical solution for digital long-time preservation. This means more than just having a permanent storage for documents. The other major aspect is to keep the files useable.

The OAIS (Open Archival Information System) reference model describes, beside software, strategies and workflows for such projects.

The key to accessibility of digital files is format migration.

Long-time preservation means to find ways and mainly file formats, which are considered to be capable of for this task. In the last years the archival community found a series of those formats, like uncompressed TIFF or JPG2000 for images or PDF/A for textual documents.

Nobody knows if and how future programs will be able to interpret nowadays digital files, therefore it is important to keep them as simple as possible. Please note, that this is by far not the final and probably most optimal solution, e.g. TIFF-files take a lot of memory, and the discourse is still ongoing. Furthermore, not for every type of digital file do such long-term formats exist. E.g. for movie files there is only a proposal for a file format yet, which is still being discussed.

At this point the idea of so-called preservation action comes into play, the migration of file formats to long-term-stable formats. The main advantage of this approach is to minimize the amount of file formats in the repository, while keeping as much information as possible. This is achieved by an automatic format migration right after the actual ingest. The benefit is to keep file formats useable, which are no longer supported

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 9, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2019. ISSN: 1314-4006, eISSN: 2535-0366

by standard software and be best prepared for future software versions and their requested file formats.

2 The Open Archive Information System Reference Model - OAIS

When dealing with digital documents in an archive, the Open Archive Information System reference model is the most important standard model. The aim of the model is to offer a solution to prevent the uselessness of all kinds of digital documents. There can be many reasons for this: For example, there could be no more supporting hard- or software for the archived file format. Or it may be a very specific file format, which is used just by one company. Or an important information like the layout gets lost when it is saved separately from the context file.

OAIS was originally initiated and developed by the NASA (National Aeronautic and Space Administration) with support of the ESA (European Space Agency) in the 1990's. The first version was presented in May 1999 by the Consultative Committee for Space Data Systems (CCSDS) and got evolved to an ISO standard in 2003. The current version was released in June 2012 and transferred to an ISO standard just two months later[1] (Schrimpf, 2014, pp. 7-14).

In the community of archives and associated institutions, OAIS got well-received and is now widely used in archives. It should be noticed at this point, that the word "open" in OAIS means open source and development of the *model*, it does not aim to be an open access archive without any terms of protection. You may have noticed, that the fact of the open developing of the model indicates changes and no final standards or solutions; it is per definition a work in progress.

OAIS is all about information packages, which contain the digital content itself as well as information about how the package arrived at the repository, how and by which tools it was treated and processed. The OAIS distinguishes between three different types of packages: The submission information package (SIP), the archival information package (AIP) and the dissemination information package (DIP).

The starting point of the OAIS-workflow is a SIP, an information package that is to be archived (see Fig.1). Beside the SIP itself, more information, like a working history or the original file format, are added and the digital file gets converted if necessary. This, in combination with a list of attributes for integrity and authenticity, forms the AIP.

It might look like just a small feature, but converting the digital file into a standard OAIS-package is one of the central aspects of the model. As mentioned before, at this point the document achieves a state that is considered best for the future (from today's perspective of course). It is almost certain, that even this digital file needs to be converted every now and then in the next decades and centuries.

[1] See ISO 14721:2012 or see the original version from June 2012 as the „magenta book", which can be found for free download.

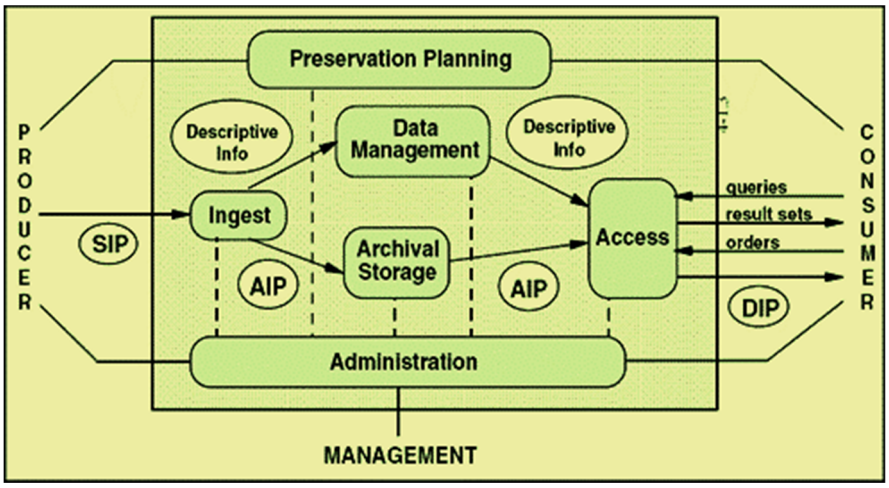

Fig. 1. OAIS functional Entities

The process of uploading new packages into a repository is called ingest. It results in having an AIP stored in the repository.

This AIP is stored in the archival storage and is handled by the data management. If a visitor of an archive or any other user requests information about or from an AIP, he or she will always be handed a DIP, which is actual a copy or duplicate of parts of the information in the AIP. How big a DIP is or what it contains depends on the permissions the requesting user has.

To emphasize this again, the AIP never gets changed, as long as the standards of the OAIS do not change. If in the future another file format than the current one is considered better (meaning: more reliable and stable for future use), a new AIP is created, derived from the original AIP. This is basically one of the main actions which are summarised as preservation actions (Schrimpf, 2014, pp. 15-23 and ISO 14721:2012).

3 Startext SORI – a Ready to Use Full OAIS Compliant Software Suite

To make the theory of the OAIS reference model useable for the usual working day in archives, startext developed an application called SORI. It is a ready-to-use software for digital long time preservation. It can be used as a standalone solution to preserve files of any age for the future. For a user it takes just two clicks to not only ingest a large number of SIPs but also to ingest whole archive stocks. After this it is possible to have full text searches over all ingested text files.

Startext SORI comes with pre-set, ready to use configurations for automatic ingest, which means that all necessary settings are complete and the processing of information packages is possible right after the installation. While running, all processes are documented twice: in the SORI-database and as a PREMIS-XML stored as part of the AIP.

The preconfigured workflow includes an antivirus scan, executed by the antivirus tool which is already used on the customer's infrastructure.

For some types of files SORI executes extra features, e.g. extract full text from text documents or, available as a beta version, a voice-to-text recognition for audio or film files with the Open Source Acoustic Models for German Distant Speech Recognition. Thumbnails are also generated for all files. Another part of the automatic process is the generation of a MD5 checksum, which is essential for detecting changes of any kind in a file. Another level of securing the files is the calculation and comparison of so-called significant properties (see chapter 5 for further discussion). These are technical properties, which are checked at critical points of the ingest or preservation and should never change.

Once stored in the backend, SORI offers multiple ways of making AIPs available and usable to/by users. For example it is possible to get access not only to the archive-package as a whole but also to single file of it, e.g. the main document without any metadata.

An additional way to keep control over the stored files is adding a digital watermark or, in the case of text documents, blackened text passages.

In our modern working world, many different file formats are used. Just think about text documents, which are mostly created as .docs, .docx, .pdf or any number of open source formats. Some of these formats can easily be converted into other formats. But most text file formats, like MS Word files, contain dynamic information, e.g. the date, and this is a critical point. That is because in those cases we must distinguish between the template and the printed document. The digital document is in fact just a template, the printed version is subject to change, depending on the time and circumstances it is printed.

In order to make that clear let us take a look at a real-life example: A letter written with Microsoft Word uses the dynamic date feature, so the actual date will be added automatically. If the letter is printed and sent, it is dated with today's date. Two weeks later you decide to archive the digital document because the printed version was sent. When SORI converts the MS Word file to PDF/A it works the same as if the document is printed, which means it is archived with the date of the file conversion and not the actual date of shipment as a physical object. This can be problematic when you archive parts of the correspondence with a visitor at once because you maybe loose the chronological order. So, the information in the AIP is like a print of the original template. The main question, that appears here, is: Is the date a significant property of the digital document or not?

In this case this might be a minor question. But what about Microsoft Excel files with many dynamic cells? What if not the resulting values but the dynamic itself is significant? Currently there is no way to keep those dynamics in the archived file version.

Another example are text converters in general: Some of those PDF/A converters just convert the content of a file but do not preserve the layout, so it is changed or completely destroyed.

4 Significant Properties and File Formats for the Future – Subjects to Change?

The current subject of discussion is the significant properties of digital files. Which properties of a digital file are considered to be significant varies from institution to institution, or even from file to file. Even if the presentation of a document seems to be the same in two different file formats, e.g. .doc and PDF/A, on the level of source code it is impossible to keep every byte the same. This has been shown with the example of the MS Word letter.

To make sure, that the AIPs contains the correct information and keeps the document in its original form, an automatic check of significant properties is necessary. Those significant properties of the digital file must never be changed.

As a start, SORI nowadays treats the number of pages of text documents and the width and heights of images as significant properties. In the workflow this means SORI retrieves this information of a file, then converts the file to the matching archival file format and compares the retrieved significant properties to assure that the file conversion did not do any damage.

This is only the starting point of verification of files for the long-term preservation. Other characteristics, e.g. the number of characters or even full text, are about to get added to the checking process.

As already explained, what parts and information of a document are most important varies from document to document, even if the file format is the same. Today researchers and the archive community are asked to specify standards for significant properties, which should be included into file format migration checking procedures.

The already mentioned issue of dynamic parts of digital documents is very important here. When we return to the MS Excel file mentioned earlier containing a lot of dynamic cells, the immutability of significant properties is a contradict to the dynamics. One could argue, that the significant property is just the values shown in the cells. But it is also possible, and even more likely, that it is not the values but the dynamics itself, which is more important for future users. There is no clear solution for this problem yet. For some MS Excel files the dynamics are not relevant yet for others they are.

Besides, disregarding the issue of significant properties, some of the file formats widely accepted as suitable for long-term preservation cannot preserve the dynamic parts of a document

It even gets worse. One could take the idea of significant properties to the next level and encounter what we call blurred significant properties. Those would be properties with a range of inaccuracy accepted. The main question by defining such properties is what level of similarity is required to accept a digital file as still being intact after a format migration. Who will specify such level and set some standards?

As mentioned in the beginning of this paper, all set standards and definitions are first attempts to handle the massive number of files of all kinds, which are subject to archiving in the next years. Hence solutions that refer to those standards and definitions can never be called final and always must be able to adapt to those changes in research and the archive community. Conversely this means, that the currently recommended and/or popular file formats used as long-term file formats are not the ultimate answer. For

example, the formats of .tiff for images or .mp4 for movie files are more or less just container formats.

In terms of movie files this means there is no specific codec used on the binary level. Instead there are countless codices in use. While those codices are still present and can be interpreted by modern software there is no problem. But it is very likely that one day a codec gets too old to be recognized and inaccuracies will happen by playing the file or rather it's generated DIP. This is one point, where companies and research institutions have to improve the OAIS related recommendations.

5 Conclusion

The two main issues of digital long-term preservation today are: what parts and in which formats digital documents of any kind can be archived indefinitely. The Open Archive Information System reference model offers a lot of first answers to those questions and startext SORI puts them into a ready-to-use software. With this program long-term preservation of digital documents becomes possible for archives of all size. But there are still lots of problems to be solved and further discussions and modifications are necessary.

Finding and agreeing on appropriate long-term file formats and finding and agreeing upon standards for significant properties for all relevant files types remains an important and urgent task.

As long as those questions are not answered for every file format, the old Nintendo quit-screen is still right: "Everything not saved will be lost".

References

Deutsches Institut für Normierung e.V. (2012). *DIN ISO 14721 Open Archive Information System (OAIS) Reference Model* (ISO 14721:2012)

Keitel, C. (2018). *Zwölf Wege in das Archiv – Umrisse einer offenen und praktischen Archivwissenschaft.* Stuttgart

Language Technology Group (n. d.). Open Source Acoustic Models for German Distant Speech Recognition. Retrived March 5, 2019 from: https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/acoustic-models.html

Neuroth, H. & Oswald, A. et al (Eds) (2009). *nestor Handbuch - Eine kleine Enzyklopädie der digitalen Langzeitarchivierung Version 2.0.* Glücksstadt

Schrimpf, S. (2014). *Das OAIS-Modell für die Langzeitarchivierung - Anwendung der ISO 14721 in Bibliotheken und Archiven.* Berlin, Vienna & Zürich, DIN Deutsches Institut für Normung e.V.

Received: June 05, 2019
Reviewed: June 21, 2019
Finally Accepted: July 01, 2019