**iPRES 2024 • iPRES 2024 Proceedings**

# Addressing The Problem of File Formats Obsolescence: Italian Guidelines on File Format Conversion for the Long-Term Preservation of Electronic Records

**Stefano Allegrezza**[1]

[1]**University of Bologna**

**Published on:** Aug 30, 2024

**URL:** https://ipres2024.pubpub.org/pub/oswmkgvc

*Abstract* – The issue of file formats conversion is crucial in the field of digital repositories. In fact, electronic records are encoded in the most diverse file formats (text, still images, audio and video recordings, technical records and many other digital media formats), but they are bound to become rapidly obsolete (and in some cases already are) and need to be converted from one file format to another in order to preserve the authenticity, reliability, integrity and usability of such records as evidence of business functions, processes, activities and transactions.

Consequently, the challenge of identifying when file formats need to be converted, which formats to choose for conversion, and which approaches to apply from an operational point of view begins to emerge.

However, despite being theorized as one of the most successful digital preservation solutions, format conversion has not yet been sufficiently practiced. Even when the need to transfer electronic records from an obsolete format to a more modern one is adequately felt, the process is often delayed due to a lack of knowledge and expertise, as well as a lack of clear and precise guidelines and advice on how to carry out the conversion, including technical and operational methods.

This paper aims to address this issue by referring both to the international standards and guidelines and to the Italian "Guidelines on the creation, management and preservation of electronic records", issued by the Italian Agency for Digital Government. In particular, it will focus on the operational methodology for file format conversion contained in its Annex 2 "File formats and conversion".

*Keywords* – Digital repositories, File formats, Obsolescence, Conversion, Migration.

*This paper was submitted for the iPRES2024 conference on March 17, 2024 and reviewed by Kate Murray, Jan Hutar, Tobias Steinke and Sam Alloing. The paper was accepted with reviewer suggestions on May 6, 2024 by co-chairs Heather Moulaison-Sandy (University of Missouri), Jean-Yves Le Meur (CERN) and Julie M. Birkholz (Ghent University & KBR) on behalf of the iPRES2024 Program Committee.*

# Introduction

In the field of digital preservation, the issue of file formats conversion is crucial. In fact, the universe of digital repositories encompasses different types, from national repositories of digitized documents that can be freely accessed, to institutional repositories set up and managed by universities to house the scientific production of researchers, to collections digitized by libraries and archives. A digital repository can include text, still images, audio, video, technical drawings, or other digital media formats. The electronic records contained within it are encoded in the most heterogeneous file formats, but almost inevitably destined to become obsolete (indeed, in some cases they already are, such as in the case of PCX or TGA image file formats, RA and RAM audio file formats, DjVu, plain text with non-UTF encodings, just to name a few). So, the problem of deciding – sometimes with some urgency – when file format conversion is required, which formats to select for

conversion, and which methodologies to use from an operational point of view begins to arise. However, to date, format conversion, although theorized as one of the most effective digital preservation strategies, has not yet been sufficiently practiced, at least in most situations. In many cases, even though the need to transfer electronic records from an obsolete format to more up-to-date formats has been felt, the process has not yet been started because there is both a lack of knowledge and expertise, and a lack of clear and precise guidelines and advice about how the conversion should be carried out, including technical-operational methods.

## Terminological Issues

Before going any further, it is necessary to give some terminological clarifications. The ISO 13008 standard [1] defines "conversion" as "the process of changing records from one format to another". According to other sources, this process is called "migration", whereas the cited standard defines "migration as the process of moving records from one hardware or software configuration to another without changing the format". In any case, in the literature [2], the term "conversion" is sometimes interchangeable with "migration"; sometimes, however, the term "migration" takes on an opposite meaning.

In Italian law, file format conversion is called "replacement transfer" (in Italian: "riversamento sostitutivo"), whereas the transfer of an electronic record from one storage system to another (without changing the file format) is called "direct transfer" (in Italian: "riversamento diretto"). In fact, the Resolution of the National Centre for Information Technology in Public Administration (in Italian: "Centro Nazionale per l'Informatica nella Pubblica Amministrazione, CNIPA) No. 11 of 19 February 2004 defined "direct transfer" as the process that transfers one or more records from one optical storage medium to another, without altering their electronic representation (i.e. the file format), whereas "replacement transfer" was defined as the process that transfers one or more records from one optical storage medium to another, altering their electronic representation (please note that at that time, optical media were considered among the best storage media). While for the first process no particular operating methods were envisaged, for the second the aforementioned Resolution established the affixing of a time stamp and a digital signature by the digital preservation officer.

## Reasons for file format conversion

The need to perform a file format conversion can be due to various reasons. According to the ISO 13008 standard [1] the reasons can be summarized as follows:

a) ***obsolescence***: the file formats of some electronic records contained in the digital repository have become obsolete and therefore a format conversion is necessary (although other strategies, such as emulation, are possible); for example, it could be that records encoded in an obsolete image format need to be converted to a more up-to-date file format.

b) ***proprietary issues***: the electronic records contained in the digital repository are encoded according to proprietary formats and must therefore be converted to non-proprietary formats, as in the case of converting

documents in DOC format (the ‘old’ Microsoft format for text documents) to PDF/A.

c) ***technological changes***: the electronic records stored in obsolete but still readable formats must be converted to current formats due to a change in the technological systems.

d) ***interoperability reasons***: electronic records are converted to a format that guarantees perfect interoperability with certain technological infrastructures.

e) ***legal reasons***: digital material needs to be converted according to explicit legal or regulatory requirements regarding formats or service providers.

When converting file formats, the final result can be one of the following [1]:

1) ***replacing one format with another***. For example, this may be due to changes in the software tools used in the digital repository, abandonment of legacy formats at risk of obsolescence, or changes in the standard format used by the digital repository for online publication.

2) ***creating an additional version*** in a different file format to meet usability requirements. For example, a report was created in a word processing format (e.g. DOCX) but needs to be converted to another format (e.g. PDF) to be published online.

In the first scenario, maintaining access to information in the digital repository means making sure that it is fully available and usable over time and through changes. In order to maintain access, it is necessary to convert file formats not only because they naturally age and can become risky but also for reasons related to technological changes (of course, alternatives are also possible, such as emulation or the development of viewers for old formats). In other words, a file format may still be current but there is a need to convert it because the technological environment used to manage the digital repository has changed. If file formats are not converted with a proactive approach, there is a risk that accessing or using the information in the required manner will become impossible or that it will be necessary to rely on specific software. Conversely, when formats are replaced, support for old file formats may be eliminated and the original files potentially deleted altogether, which entails certain risks.

In the second scenario, rather than converting an electronic record to a new format, additional versions of electronic records in different formats are created to allow new forms of access and use, such as sharing or publishing information, using information in new ways, and aggregating information from various sources. This does not imply that the original format is obsolete; rather, more than one format may be necessary to meet all requirements for the same information. However, formats should not be multiplied unnecessarily: if a single format meet all access needs, that is usually the best solution [3]. The typical example is a digital repository with a collection of images. These images are used in many different situations, and although there is a standard “package” of different versions, occasionally a situation requires the creation of a new version.

Usually, the “master” of each image is a high-resolution lossless TIFF file, which can be opened with an image viewer (e.g. the one supplied with the operating system); a series of JPG versions of each image (“derived” file formats), optimized for the human eye, are stored at different resolutions and qualities; GIF images, optimized for web, are often used as thumbnails.

## The time of File format Conversion

Once it has been established that it is necessary to convert electronic records from one file format to another, another important issue is to decide when to carry out this conversion. There are basically three strategies for converting file formats [3]: 1) on-demand conversion; 2) early conversion; 3) late conversion. The strategy chosen will largely be dictated by the motivation behind the format conversion but may also depend on the technical environment or other needs of the digital repository.

1) ***On-demand conversion***. This strategy relies on servers to perform the conversion dynamically. It means that conversion of an electronic record to another format is carried out when a request for that format is received. It generally operates on a single electronic record at a time, although batch conversions may also occur on-demand. This process may be automated or may require an individual to manually convert electronic records on-demand. This strategy can be applied to replace formats but is most often applied to create additional versions of electronic records in different formats as required [3]. For example, the digital repository might offer users electronic records in different formats (e.g. PDF, DOCX and ODT). However, it is not convenient to store each electronic record in all formats, but it is better to store only one electronic record (usually in the most complete format) and to generate the others when a request for an electronic record in a different format is received.

This strategy has many advantages: it is not necessary to store several copies of each electronic record in each format, so the storage space is reduced: only one electronic record is needed, the conversion of which is done dynamically on-demand (however, it is possible to store a converted electronic record to speed up any future requests); it is not necessary to convert a large number of electronic records at once, which may take a long time; adding new electronic records to the system can be quite simple, as it is not necessary to provide all required formats in advance; the system can be updated to provide different formats as required, again without having to process all existing electronic records in advance.

Anyway, there are downsides as well. It is not so simple to verify the quality of converted files, even if QA/QC workflows that can be implemented to measure quality and conformance. If this strategy is adopted, it is necessary to ensure that the conversion process is sufficiently reliable for all the needs. The digital repository may not allow dynamic requests for electronic records in different formats. For instance, if electronic records are accessed via a network share, there is no way to act on an on-demand conversion server. On-demand conversion can be slow or overly burdensome for systems, depending on the size, complexity and number of conversions. This strategy generally only makes sense for static information. If it is necessary for users to

modify data, an on-demand format conversion strategy may not work unless there is a clear master version, which can only be modified in that version.

2) ***Early conversion***. This strategy means converting an electronic record into a different file format as soon as possible (but not on demand). Early conversion is a batch processing strategy of converting a set of electronic records in one common format to another that best suits the digital repository needs and is generally a replacement process. For example, if the digital repository has decided to use a new format provided by upgraded software, all previous electronic records are converted to the new format.

This strategy has many benefits. The number of different file formats to support is greatly reduced by converting electronic records onto a standardized set of formats. This can mean that information is always encoded in the currently supported formats, so support, maintenance and software licensing costs are reduced; the risk of electronic records' format obsolescence becomes negligible. It is possible to review information and ensure quality. With frequent conversion, these processes are streamlined and each conversion benefits from previous experience [3].

Obviously, that are disadvantages as well. Each electronic record must be converted more frequently, and each conversion has an associated cost and risk of information loss. If the original or new formats are fairly recent, conversion tools may not be as readily available, may have bugs, or may not handle complex or unusual electronic records well. This, too, can affect the cost and quality of the conversion process. The new format may not be as widely supported, so it may be necessary to create other formats if one wants to share information with users who have not yet upgraded. If it is necessary for the same information to be accessible in several formats, storing all converted electronic records will require more space than on-demand conversion [3].

3) ***Late conversion***. This means that the digital repository has decided to postpone the conversion to a later stage in the life cycle and workflow, sometimes until the last useful moment. Of course, the definition of "last useful moment" varies greatly depending on the digital repository's risk/benefit assessment. For instance, after a risk assessment of the electronic document formats used in the digital repository, one may find that one has a large amount of legacy information stored in a dozen different file formats, some of which is no longer accessible with the current software. Some of this information may no longer be needed for active digital repository's needs, which is why a preservation strategy is employed. However, some information is still occasionally needed, so a different file format is chosen.

This strategy has many advantages. Each electronic record is converted less frequently, so there is less risk of information loss and overall costs are lower. If the target format is widely adopted, more conversion tools will probably be available to use, and existing conversion tools probably will be able to handle unusual or complex files better because there has been time to fix bugs and edge cases. It may be possible to discard older information that is no longer useful for the digital repository, thus avoiding the need to convert it.

There are disadvantages, too. The digital repository will have a greater variety of formats in use at any given time. This can increase software support, maintenance, and licensing costs; reduce flexibility in choosing different software; and prevent older information from being usable in newer contexts. More electronic records will probably need to be converted and a greater variety of file formats at once, making the project larger to manage and more complicated to evaluate from a quality point of view. If the same information need to be accessible in multiple formats, storing all converted electronic records will require more space than on-demand conversion. Finally, one could make a mistake in assessing the "last minute" and find that converting some information is no longer economically or technically feasible.

Early and late conversions are completely different from on-demand conversion and are just variations of batch conversion processes, but with different risks and costs due to the timing of the conversion. These strategies are often confused with each other and there is a continuum of combinations between the two strategies; the extreme ends are explained to demonstrate the different trade-offs involved. There is no one-size-fits-all strategy, and each has pros and cons: only by assessing the needs of the digital repository can the right balance of risks, costs and benefits be determined.

# The Italian Guidelines on File Format Conversion

## The Guidelines

In the Italian context, the main reference to the topic of format conversion is Annex 2, "File formats and conversion" to the "Guidelines on the creation, management and preservation of electronic records" published by the Italian Agency for Digital Government (AgID) which became mandatory as of January 1, 2022 [4].

According to these Guidelines, the conversion of records from the original format to another may take place several times and at different times, both for management and preservation purposes. In the second case, file formats may need to be migrated when they begin to be obsolete. The tool for assessing the level of obsolescence of electronic records is the so-called "interoperability assessment", which all public administrations and private companies should carry out at least every year. The Guidelines propose a quantitative methodology for evaluating file formats that consists in evaluating a group of nine factors (Fig. 1) each of which is assigned a numerical value (score).

- Standardization (from 0 to 3 points)
- Disclosure (from 0 to 3 points)
- Proprietary (from 0 to 4 points)
- Extensibility (from 0 to 2 points)
- Level of metadata (from 0 to 3 points)
- Robustness (from 0 to 2 points)
- Device independence (from 0 to 4 points

There are also two other evaluable factors (to which, however, Annex 2 does not assign a specific score):

- Backward and forward compatibility
- Textual or binary encoding

The sum of these values is called “interoperability index” that can vary between a minimum of 0 (zero) and a maximum of 21. A value equal to 12 is considered as a sufficiency threshold. File formats that reach an interoperability index equal to or greater than 12 are “acceptable” while lower values reveal objective problems that must be addressed as soon as possible using, for example, conversion processes or other methodologies.

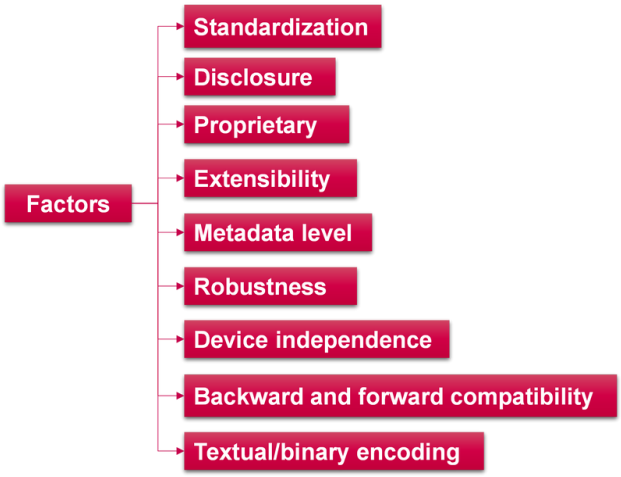


Figure 1 The factors to evaluate when calculating the “interoperability index”.

The tool for assessing the level of obsolescence of electronic records is the so-called “interoperability assessment”, which all public administrations must carry out every year. Following the interoperability assessment, consideration should be given to planning for the conversion of electronic records from a file format to another in order to ensure their preservation over time. The conversion process should be carried out according to the guidance provided in the previously mentioned Annex 2.

In addition, when choosing new file formats, it is important to consider the technical peculiarities of the source and the target file format, with particular reference to both the loss of data and metadata and the different quality or technical representation of them, stating that, in some cases, preservation of the electronic record in its original format is mandatory.

In this regard, it is important to consider the projects that have already identified the essential characteristics that need to be retained in a file format conversion process. Certainly, the most relevant in this area is the InSPECT (Investigating the Significant Properties of Electronic Content Over Time) project that was funded by JISC between March 2007 and March 2009 under the Repositories and Preservation program [5]. It was led by the Arts and Humanities Data Service (AHDS) Executive until its demise in March 2008, and then by the

Centre for e-Research (CeRch) at Kings College London. CeRch was assisted by project partners at The National Archives (TNA). The purpose of INSPECT was to establish a generalized methodology for assigning certain standard "significant properties" to certain categories of electronic records. By "significant properties" the project initiators meant certain characteristics of electronic records that need to be preserved over time. For example, some of these properties are the content of the records, the metadata that contextualizes their production and function, their appearance (e.g., layout, colors, etc.) [6], the purpose for which they were produced, or even their logical structuring. According to the project initiators, deciding which of these properties should be standardized and considered in electronic records preservation practices is a key prerequisite for their improvement.

## Basic principles

Annex 2 specifies in detail the principles on which the conversion process should be based. These principles can be summarized as follows:

1) When the conversion is carried out for file preservation purposes, a copy of the file in its original format may and, in certain cases provided for by law, must also be preserved.

2) When a conversion is made for preservation purposes, the converted file is a digital copy of an electronic record and, as such, the conformity of the copy is certified in accordance with current legislation, including the digital signature of a notary or public officer according to the "process certification" referred to in Appendix 3 of these Guidelines.

3) It is important that in the choice of new file formats (and, consequently, of containers, file packages, digital streams and codecs), as well as in the methodological choice regarding the execution of the conversion, the technical peculiarities of the source format and of the transferred format are taken into account, with particular reference both to the loss of data and metadata and to the different quality or technical representation of the memories.

4) Conversion must always take place to formats that improve interoperability, or at least do not worsen it (as, for example, established by calculating the interoperability index). Conversions are especially not appropriate:

- from an open format to a closed format
- from a non-proprietary format to a proprietary format
- from a device-independent format to a device-dependent format
- from a format with embedded metadata to a format without embedded metadata

## Technical-operational procedure

Annex 2 describes the technical-operational procedures for performing a bulk conversion for the purpose of electronic records preservation. The conversion process shall include the following steps:

1. Each conversion procedure must be accurately described.

2. The conversion of a file format must be performed by a “certified process” that guarantees the integrity of the result and the reproducibility of the process itself, also to fulfil legal requirements.

3. In order to ensure that the converted format conforms to the original format, for each converted file of a bulk conversion the process must produce a proof of the specific conversion of that file. The proofs about records converted as part of the same conversion process are collected in a “conversion register”. This register must contain:

- a time reference enforceable against third parties (start and end date and time of the procedure)
- information on the information system used
- the name of the file
- the location of the file in the file system
- the external and internal metadata
- the file format and its version
- any technical information about errors, anomalies or ambiguities found during conversion.

4. In compliance with applicable privacy laws, conversion to another file format constitutes a further opportunity to comply with the obligations regarding the adequacy, relevance, minimization, and accuracy of the personal data contained therein, as well as the lawfulness of their processing and their possible pseudonymity.

5. Where there are legal or other constraints on preserving the bitstream, the original record must be preserved together with its conversion into a more interoperable format. This logical association must also be written in the “conversion register”.

6. Since there are format reversions that preserve the record content substantially unchanged, it is necessary that such possibilities, if they are identified, be described specifying how the record content is substantially preserved and how these cases are to be handled.

The considerations here apply to every type of file format (envelope formats, package formats, container formats, binary streams, and codecs, etc.).

## Conclusion

File format conversion is considered one of the most important preservation strategies, and many digital repositories are starting to make some conversions or are planning them for the near future. Some researchers have prototyped a set of e-services that serve as a framework for understanding content preservation, automation, and computational requirements on preservation of electronic records. This framework consists of e-services for (a) finding file format conversion software, (b) executing file format conversions using available

software, and (c) evaluating information loss across conversions [7]. Some academic institutions have developed file converters and let them free for the benefit of students, who can easily convert they office automation documents (e.g. in DOCX, PPTX o XLSX format) in PDF format needed for submission purposes. An example is the VMEG tool Kit [8]. Unfortunately, despite being one of the most widely theorized file format conversion strategy, it is not yet sufficiently practiced, at least in the Italian context. Even when the need to convert electronic records from an obsolete format to a more modern format is adequately felt, the process is often delayed due to a lack of knowledge and expertise, as well as a lack of clear and precise guidelines and advice.

The Italian Guidelines try to fill this gap, providing some useful and operational indications on how the conversion process should be managed, but are lacking on some aspects. For example, they do not provide information about the software application to be used for conversion. There are many tools for converting electronic records. Some are proprietary, some are freeware, and some are open source. However, coverage of formats can be inconsistent. For popular formats, such as images, the choices may be numerous, but for niche or older formats the choices may be very limited. For formats with poor support, it may be necessary to perform two conversions, using an intermediate file format to bridge the gap between the format in use and the desired format. In some cases, it may be necessary to commission customized software to perform the conversion, especially if file formats are themselves bespoke. It is important to assess whether these tools fully support the essential features and metadata that need to be converted, and not only whether they convert from source to target format.

In this regard, the Italian guidelines are lacking: for example, they do not define metrics to measure the loss of information after a conversion and to verify whether the features that are being tried to be preserved have survived the conversion process; or , how to use them to test the accuracy and quality of format conversions, etc.  In recent years, a number of projects have addressed the issue of quality assurance and digital preservation actions, such as the AQUA [9] [10], SPRUCE [11] and SCAPE [12] projects, but their results need to be consolidated and brought together in a common vision. It would be highly desirable for the digital preservation community to begin to seriously consider these issues and to develop guidelines and recommendations to provide internationally shared guidance and advice on how to deal with the problem of file format conversion.

## References

1. ISO 13008 Information and documentation — Digital records conversion and migration process. ↩
2. Digital Preservation Coalition (DPC), Digital Preservation Handbook, https://www.dpconline.org/handbook/organisational-activities/preservation-action. ↩
3. The National Archives of United Kingdom, File Format Conversion, https://cdn.nationalarchives.gov.uk/documents/information-management/format-conversion.pdf. ↩

4. Italian Agency for Digital Government, Guidelines on the creation, management and preservation of electronic documents, Annex 2 “File formats and conversion”, https://www.agid.gov.it/sites/default/files/repository_files/allegato_2_formati_di_file_e_riversamento.pdf. ↩

5. InSPECT (Investigating the Significant Properties of Electronic Content Over Time) project, https://significantproperties.kdl.kcl.ac.uk. ↩

6. Archives of New Zealand, File format migration, https://www.archives.govt.nz/ manage-information/how-to-manage-your-information/digital/file-format-migration. ↩

7. P. Bajcsy, R. Kooper, L. Marini, et al. A framework for understanding file format conversions, in Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop, March 2010, https://dl.acm.org/doi/10.1145/2039274.2039284. ↩

8. D. Bhargav Reddy, C. Lakshma Reddy; S. Pulluri, et al., VMEG Mini Tool Kit – An Intelligent Approach for File Conversion, https://ijirt.org/master/publishedpaper/ IJIRT154032_PAPER.pdf. ↩

9. P. Wheatley, B. Middleton, J. Double. People Mashing: Agile Digital Preservation and the AQuA Project, https://services.phaidra.univie.ac.at/ api/object/o:294255/download. ↩

10. AQUA (Automating Quality Assurance) Project, https://web.archive.org/web/20220623161628/https://wiki.opf-labs.org/display/AQuA/Home ↩

11. SPRUCE (Sustainable PReservation Using Community Engagement) Project, https://web.archive.org/web/20130716070354/https://wiki.opf-labs.org/display/SPR/Home. ↩

12. SCAPE (Scalable Preservation Environments) Project, https://scape-project.eu. ↩