File formats conversion issues in digital libraries

Stefano Allegrezza[1]

[1] *University of Bologna, Italy*

Abstract

The issue of file format conversion is critical in the realm of digital libraries. In fact, a digital library can include text, still images, audio, video, digital documents, or other digital media formats. These digital objects are encoded in the most diverse file formats, but they are eventually doomed to become obsolete (and in some cases have already done so). As a result, the challenge of identifying when file format conversion is required, which formats to choose for conversion, and which approaches to apply from an operational standpoint begins to emerge. However, while being theorized as one of the most successful digital preservation solutions, format conversion has yet to be sufficiently practiced. Even when the need to transfer digital objects 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 on how the conversion should be carried out, including technical-operational methods. This paper attempts to address this problem by emphasizing the rationale for format conversion (Why), the circumstances under which it must occur (When), and the methods by which it can be done (How), citing both international norms and guidelines and the Italian "Guidelines on the creation, management, and preservation of electronic documents," published by the Italian Agency for Digital Government.

Keywords

Digital Library, File formats, Migration, Conversion, Obsolescence,

1. Introduction

In the field of digital libraries the issue of conversion of file formats is very important. In fact, the universe of digital libraries encompasses different types, from national repositories of digitised documents that can be freely accessed, to institutional repositories set up and managed by universities to house the scientific production of researchers, to digitised library collections by individual libraries and archival institutions [1]. A digital library «can include text, still images, audio, video, digital documents, or other digital media formats or a library accessible through the internet» [2]. The digital objects contained within them are encoded in the most heterogeneous file formats, but inevitably destined to become obsolete (indeed, in some cases they already are, such as in the case of of RTF, DjVu, plain text with non-UTF encodings, PCX or TGA image file formats, 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 theorised as one of the most effective digital preservation strategies, has not yet been sufficiently practised, at least in most situations. In many cases, even though the need to transfer digital objects 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 advices about how the conversion should be carried out, including technical-operational methods.

This paper aims to address this issue by highlighting the reasons behind format conversion (Why), the time when such conversion must be carried out (When) and the methodologies for implementing it (How), referring both to the international standards and guidelines and on the

Italian "Guidelines on the creation, management and storage of electronic documents", issued by the Italina Agency for Digital Government (in particular with reference to Annex 2 "File formats and conversion") [3]. It also proposes a methodology for performing the file format conversion.

2. Definitions

Before moving on, it is necessary to make a clarification of terminology. The ISO 13008 standard [2] defines "conversion" as "the process of changing records from one format to another". According to other sources this process is called "migration", while the above-mentioned standard defines "migration the process of moving records from one hardware or software configuration to another without changing the format". Anyway, in the literature [4], the term "conversion" is sometimes interchangeable with "migration"; sometimes, on the contrary, the term "migration" takes an opposite meaning.

In Italian law, file format conversion is called "replacement transfer" ("riversamento sostitutivo") while the transfer of a digital object from one storage system to another (without changing the file format) is called "direct transfer2 ("riversamento diretto"). In fact, the Resolution of the National Centre for Information Technology in Public Administration (CNIPA) no. 11 of 19 February 2004 defined "direct transfer" as the process that transfers one or more stored documents from one optical storage medium to another, without altering their computer representation (i.e. the file format), whereas "replacement transfer" is the process that transfers one or more stored documents from one optical storage medium to another, altering their computer representation.

While for the former process no particular operating methods were envisaged, for the latter the methods described in Article 3, co. 2 and Article 4, co 4 of the cited Resolution – which concerned, respectively, the replacement preservation of electronic documents and that of analogue documents – the affixing of the time stamp and the digital signature by the person responsible for storage was planned.

3. Why to convert file formats

When the issue of format conversion is discussed, the first thing to point out is *why* it is necessary to convert file formats. The need to perform a file format conversion can be due to various reasons. According to the ISO 13008 standard [5] the reasons can be summarised as follows;

a) *obsolescence*: the file formats of some digital objects contained in the digital library have become obsolete and therefore a format conversion is necessary; for example, it could be that objects encoded in an obsolete image format need to be converted to a more up-to-date file format.

b) *proprietary issues*: the digital objects contained in the digital library 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 digital objects stored in obsolete but still readable formats must be converted to current formats due to a change in the technological systems.

d) *interoperability reasons*: digital objects 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 [5]:

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

In the first scenario, maintaining access to information in the digital library 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. 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 library has changed.

If file formats are not converted with a pro-active approach, you may find that you are no longer able to access or use your information in the way you need it, or that you are forced to use particular software. However, when replacing formats, you may find that you eliminate support for old file formats and potentially delete your original files altogether, which entails risks.

In the second scenario, rather than converting a digital object to a new format, additional versions of digital objects 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 of your access needs, that is usually the best solution [6].

The typical example is a digital library with a collection of images. These images are used in a large number of 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 formats), optimised for the human eye, are stored at different resolutions and qualities; GIF images, optimised for web, are often used as thumbnails.

4. When to convert file formats

Once it has been established that it is necessary to convert digital objects 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 [6]: 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 library.

1) *On-demand conversion*. This strategy relies on servers to perform the conversion dynamically. It means that conversion of a digital object to another format is carried out when a request for that format is received. It generally operates on a single digital object at a time, although batch conversions may also occur on-demand. This process may be automated or may require an individual to manually convert digital objects on-demand. This strategy can be applied to replace formats, but is most often applied to create additional versions of digital objects in different formats as required.

For example, the digital library might offer users digital objects in different formats (e.g. PDF, DOCX and ODT). However, it is not convenient to store each digital object in all formats, but it is better to store only one digital object (usually in the most complete format) and to generate the others when a request for a digital object in a different format is received. This strategy has many advantages: it is not necessary to store several copies of each digital object in each format, so the storage space is reduced: only one digital object is needed, the conversion of which is done dynamically on-demand (however, it is possible to store a converted digital object to speed up

any future requests); it is not necessary to convert a large number of digital objects at once, which may take a long time; adding new digital objects to the system is 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 digital objects in advance. Anyway, there are downsides as well. There is almost no way of guaranteeing the quality of converted files; if you adopt this strategy, you must ensure that the conversion process is reliable enough for your needs. The digital library may not allow dynamic requests for digital objects in different formats. For instance, if digital objects 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 on your 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 that you have decided to convert digital objects to different formats as soon as possible (but not on-demand). Early conversion is a batch processing strategy of converting a set of digital objects in one common format to another that best suits the digital library needs and is generally a replacement process. For example, if the digital library manager has decided to use a new format provided by upgraded software, you can convert all previous digital objects to the new format.

The strategy has many benefits. The number of different file formats to support is greatly reduced by converting digital objects 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 digital object format obsolescence becomes negligible. You have the opportunity to review information and ensure quality. With frequent conversion, these processes are streamlined and each conversion benefits from previous experience [6].

Obviously that are disadvantages as well. Each digital object 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 digital objects well. This, too, can affect the cost and quality of the conversion process. The new format may not be as widely supported, so you may have to create other formats if you need to share information with users who have not yet upgraded. If you need the same information to be accessible in multiple formats, storing all converted digital objects will require more space than on-demand conversion [6].

3) *Late conversion*. This means that the digital library manager has decided to postpone the conversion until the last useful moment. Of course, the definition of "last useful moment" varies greatly depending on the digital library's risk/benefit assessment.

For instance, after a risk assessment of the digital objects' formats used in the digital library, you may find that you have 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 library'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 digital object 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. You may be able to discard older information that is no longer useful to the digital library, thus avoiding the need to convert it.

There are disadvantages, too. The digital library 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. You will probably have to convert more digital objects 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 you need the same information to be accessible in multiple formats, storing all converted digital objects will require more space than on-demand conversion. Finally, you may be wrong on the side of evaluating "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 really 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 library can the right balance of risks, costs and benefits be determined.

5. How to convert file formats

According to the ISO 13008 standard [5], the conversion procedure consists of four key steps: a) planning, b) testing, c) conversion d) validation (see Figure 1).
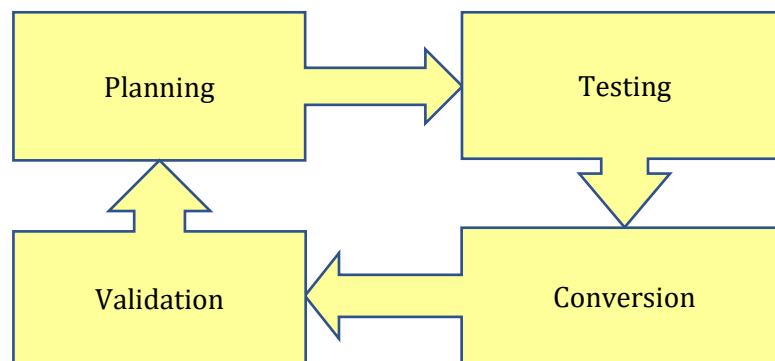
Figure 1. The four key steps of a conversion process

a) *Planning*. Is the most important step, because the greatest probability of success of the conversion process derives from a careful and accurate planning.

First, it is necessary to assess the digital objects you need to convert, that is, how one should be able to find, open, work with, understand, and trust them. These requirements may not be immediately obvious, and you should work with the manager and the primary users of the digital library to ensure that all their requirements are met. This will help to understand whether the digital objects in the formats from which you are migrating has particular characteristics that you want to ensure remain unchanged. Some conversion processes only change the file format of the digital objects, but many conversion processes also alter some characteristics. In general, very simple types of digital objects can survive a conversion process without change, but complex objects will be altered in some way. For example, you may be converting from one document format to another (i.e. from DOCX to ODT). It is possible that while the text of the document remains unchanged, the layout, colors, styles, and fonts used will be altered during conversion [5], even if in some cases the difference can be very slight. For example, the pagination of a document may change, disrupting page references embedded in the document and rendering a contract unusable. Or, if the colors of a document are considered not important in the conversion process, it is possible that users have annotated minutes using green to indicate things that are completed and red for things that are not yet completed. It is important to review the information to determine whether changes to an aspect of the information may subtly affect its meaning.

It is important to keep in mind that any conversion process potentially exposes digital objects to risk of information loss. Thus, before conversion, it is necessary to identify the key characteristics of the information that must survive conversion with no (or little) change. You should be aware that characteristics you do not consider essential may actually be essential because of the way they have been used. Some less obvious characteristics, usually related to complex or hidden functionality of the format, must also often be considered. The following is a non-exhaustive list of some of them:

- *Embedded metadata*. Many formats allow various descriptive metadata to be embedded. For example, documents that report the author of the document and photographs that report the geographic location where they were taken and the camera settings (e.g. EXIF metadata in the case of JPG pictures). Therefore, it is necessary to consider whether embedded metadata is needed in the converted files and whether the conversion tools are capable of transferring this information.

- *Embedded objects*. Many complex formats allow other files in various formats to be embedded within them. For example, documents may contain embedded images or spreadsheets, or presentations may contain video. Not all conversion tools can handle all types of embedded objects. Therefore, it is necessary to test files with embedded objects to ensure the quality of the conversion process.

- *Scripts and macros*. Some formats may contain mini programming languages. For example, documents often have a macro function to automate common tasks. In general, scripts and macros do not survive conversion processes unless the conversion is from one version of a format to another version of the same format (e.g. from Microsoft DOC to DOCX). Occasionally, another format will provide the same or equivalent support for the same embedded scripts or macros, but this is rare. Therefore, if you need script and macro support in your digital objects, you may have to manually rewrite them for the new format.

- *Digital signatures*. Some digital objects allow digital signatures to be embedded within them (or to have digital signatures in external systems related to those digital objects). Digital signatures validate that a file has been signed by an authorized user, using strong encryption on all information in the file to prove the claim. Therefore, any converted file, being different from the original, will lose this digital signature (or the signature will no longer be valid) and it will be necessary to produce a new digital signature.

It is important to ensure that the new file formats support the characteristics that are considered important (the so called "significant properties", see below) and that the conversion process maintains these characteristics during the conversion. If the new format does not support the required features, you may need to reevaluate your choice of file format or whether to migrate.

In this regard, it is important to take into account 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 programme [7]. 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 digital records. By "significant properties" the project initiators mean certain characteristics of digital records that need to be preserved over time. For example, some of these properties are the content of the documents, the metadata that contextualizes their production and function, their appearance (e.g., layout, colors, etc.) [8], 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 taken into account in digital document preservation practices is a key prerequisite for their improvement.

An important step in the planning phase is the selection of the conversion tool. There are a large number of tools for converting digital objects from one format to another. Some are proprietary, some are freeware, and some are open-source. However, coverage of formats can be patchy. 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, you may need 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, you may have to commission customized software to perform the conversion, especially if your file formats are themselves bespoke. Only by evaluating your information, your formats and your environment can you decide which tools may be right for you. It is important to assess whether the tool fully supports the essential characteristics and metadata you are trying to convert, and not just whether it converts from the source format to the target format. Complex formats can often be written in different ways, which can affect whether the characteristics you are trying to preserve survive the conversion process.

b) *Testing*. Once conversion tools have been identified and before starting the conversion process in its entirely, it is a good idea to carry out a testing phase to verify that the process is successful and that the significant properties of the digital objects and their metadata have been accurately converted. The test can be carried out on a representative sample of the digital objects to ensure that the conversion tools can perform with an acceptable level of quality. The purpose of this step is to verify that the process can transforms the existing file formats into the target file formats without loss of data and metadata, and without loss of authenticity, reliability, integrity and usability. Testing requires accurate knowledge of the existing file formats and the target file formats and specific knowledge of the initial hardware or software configuration and the target hardware or software configuration. You need a means of confirming the degree of success in order to perform automatic or semi-automatic quality checking. In addition, you need to select metadata extraction tools as well, and test them on the original and converted digital objects. You may need to use different metadata extraction tools for the source and target formats and convert their results to a common form to facilitate comparison. You will probably need to create custom software or scripts to perform this process, which can make it cost-prohibitive. However, there may be situations where converting a large amount of essential information makes it essential for assurance purposes. Recent work touching on quality assurance and digital preservation actions includes the work of the AQUA [9] [10], SPRUCE [11], and SCAPE [12]projects

c) *Conversion*. Once you have gained an understanding of your information and environment, and selected your formats and tools, you are ready to begin converting the digital library's digital objects. Some conversion tools are intended for single-file conversion, meaning that they literally take one digital object and produce another one in the new format. If you want to convert multiple digital objects, you must use a tool that allows you to convert entire folders or subfolders, or write scripts to automate batch processing of multiple digital objects. There are also companies that perform file format conversion as a service, using their own tools to perform the conversion and managing the process for you [5].

d) *Validation*. Even after initial testing of the tool, unless you choose to use an on-demand conversion strategy, additional quality assurance processes need to be established to ensure that essential information is not lost during conversion. Quality assurance criteria must be agreed in advance with the digital library manager and the primary users. You must assess whether the properties and characteristics that were to remain the same have remained the same as expected during the testing phase. Manually checking each digital object can be impractical, so the best solution is to do a spot check on a representative sample of converted digital objects, opening both the original and converted digital objects for direct comparison. This kind of direct comparison can be of various kinds. For example, while a visual-level analysis will sufficient for images, for sound objects it will be necessary to listen to the original digital object and the transferred one to understand whether there are any differences. Of course, automatic or even artificial intelligence-based comparison tools can be used.

End users of the information should be involved in this process, as they may detect subtle problems that non-users would not notice. Using metadata extraction tools, it is possible to compare the before and after values of the original and converted digital objects to be certain that the conversion process was successful. For example, one can extract the size of an image and verify that it is the same before and after conversion, or examine the audio length (not file size) of converted audio files, or compare the embedded metadata (e.g., author) of a document.

Even if the converted digital objects are intended to completely replace the originals, it is a good idea to keep the originals for a defined period of time to avoid accidental loss of functionality or information. This is true even if quality assurance processes have demonstrated that the conversion is fully successful. This serves to reduce the risk that, after the conversion process, previously unknown requirements may emerge that dictate the preservation of various characteristics of the information. This must be balanced against the competing requirements to reduce the amount of data stored.

It is often difficult to determine how long the originals should be retained, as this will depend on the importance of the digital objects, the reasons for the conversion, the organization's risk tolerance, its confidence in the conversion process, and the costs of retaining the originals and maintaining links between them and the converted digital objects.

6. The Italian Context

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 storage of electronic documents" published by the Italian Agency for Digital Government (AgID) which became mandatory as of January 1, 2022 [3]. According to these guidelines, in the context of electronic document management, activities may be required for the conversion of documents from the original file format to another one. Such conversion may occur several times and at different times for management or conservation purposes. The tool for assessing the level of obsolescence of digital objects is the so-called "interoperability assessment", which all public administrations – including those who manage digital libraries – must carry out every year. Following the interoperability assessment, consideration should be given to planning for the conversion of digital objects from one 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 take into account the technical peculiarities of the source and the target file format be considered, 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 digital object in its original format is mandatory.

Annex 2 specifies in detail the principles on which the conversion process should be based and the technical-operational procedures. Furthermore, provision is made for massive file format conversion, carried out through "certified processes" that guarantee their integrity and reproducibility. The considerations here apply to every type of file format (envelope formats, package formats, container formats, binary streams and codecs, etc.).

7. Conclusions

Format conversion is one of the most widely used preservation strategies, and most digital preservation systems contain features or system data that assume a conversion solution. 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 [13]. Some academic institutions have developed file converters and let them

free for the benefit of students, who can easily convert they office utomation documents (e.g. in DOCX, PPTX o XLSX format) fin PDF format needed for submission. An example is the VMEG tool Kit [14].

Digital libraries should also take this issue into account and begin planning conversion processes, starting with assessing the obsolescence of digital objects' formats, determining when and how to convert, defining metrics to measure any loss of information, and using them to test the accuracy and quality of format conversions. It would be highly desirable for the archival and library community to begin to seriously consider these matters and develop guidelines and suggestions to provide guidance and advice to digital libraries on how to deal with this issue.

References

[1] M. T. Biagetti, Le biblioteche digitali. Tipologie, funzionalità e modelli di sviluppo, Franco Angeli, Milano, 2013.

[2] Wikipedia, definition of digital library, URL: https://en.wikipedia.org/wiki/Digital_library.

[3] Italian Agency for Digital Government, Guidelines on the creation, management and storage of electronic documents, Annex 2 "File formats and conversion", URL: https://www.agid.gov.it/sites/default/files/repository_files. allegato_2_formati_di_file_e_riversamento.pdf.

[4] Digital Preservation Coalition (DPC), Digital Preservation Handbook, URL: https://www.dpconline.org/handbook/organisational-activities/preservation-action.

[5] ISO 13008 Information and documentation — Digital records conversion and migration process.

[6] The National Archives of United Kingdom, File Format Conversion,

[7] InSPECT (Investigating the Significant Properties of Electronic Content Over Time) project, URL: https://significantproperties.kdl.kcl.ac.uk.

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

[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, URL: https://wiki.opf-labs.org/display/AQuA/Home.

[11] SPRUCE (Sustainable PReservation Using Community Engagement) Project, URL: https://wiki.opf-labs.org/display/SPR/Home.

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

[13] 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, DOI:10.1145/2039274.2039284.

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

(All resources were last accessed on 20 December 2023).