The Italian Guidelines on Creation, Management and Preservation of Digital Records

A Proposed Methodology for File formats Assessment

Stefano Allegrezza
University of Bologna
Italy
stefano.allegrezza@unibo.it
ORCID: 0000-0002-7319-2483

Abstract – This paper aims to address the issue of file formats assessment for the preservation of digital records, which is fundamental because the chances of preserving records produced in these formats over time depend greatly on it. In particular, the paper presents and discusses the methodology proposed by the Italian Agency for Digital Government who recently published the "Guidelines on the Creation, Management and Preservation of Digital Records". This methodology is based on a quantitative method that evaluates some properties of file formats and assigns them a score; the sum of these scores is the so-called "interoperability index" that provides useful information in order to establish whether the format is prone to obsolescence or not. The methodology is explained with examples that show its potential. Some suggestions for improvement and further developments are then discussed.

Keywords – file formats, assessment, evaluation, selection, digital preservation

Conference Topics – Sub-theme2: Scanning the New Development

I. Introduction

Selecting file formats for the creation of digital records is very important since the ability to preserve over time the records depends greatly on the accurate selection of their format. For at least two decades, there has been in-depth reflection on these issues, but the guidelines and recommendations that have been published in most cases rely on a qualitative assessment [1], [2], [3], [4], [5] of file formats. Instead, a methodology based on quantitative assessments has almost never been proposed, apart from some notable exceptions such as that of the National Library of the Netherlands [6], the United States National Archives and Record Administration (NARA) [7] and the Centre de coordination pour l'archivage à long terme de documents électroniques (CECO) in Switzerland [8].

For this reason, the methodology proposed by the Italian Agency for Digital Government with the "Guidelines on Creation, Management and Preservation of Digital Records" (in Italian: Linee Guida sulla formazione, gestione e conservazione dei documenti informatici), published online on September 9, 2020 [9], is particularly interesting. These guidelines have introduced some remarkable news on the issue of file formats - such as the "interoperability assessment" which will be discussed below - in order to ensure, on the one hand, a more efficient managing od electronic records and, on the other hand, a more sustainable long-term preservation. Annex 2 to the Guidelines is entirely dedicated to the topic of file formats and migration and establishes a measured criterion, although susceptible to some degree of subjectivity, for the choice of file formats. Three chapters form it. The first, "Introduction", contains the basics and a taxonomy of file formats. The second, "File formats",

17th International Conference on Digital Preservation

iPRES 2021, Beijing,China.

DOI: 10.1145/nnnnnnn.nnnnnnn

is the main part of the Annex and describes 124 file formats divided into 16 categories:

- paginated documents;
- hypertexts;
- structured data;
- email;
- spreadsheets and multimedia presentations;
- raster images;
- vector images and digital modelling;
- fonts;
- audio and music;
- video;
- subtitles, captions and dialogues;
- containers and multimedia packages;
- compressed archives;
- administrative documents;
- applications and source code;
- cryptographic applications.

The third, "Recommendations on file formats", specifies how to make the interoperability assessment, the calculation of the interoperability index and the migration of file formats.

II. The Interoperability Assessment

The Guidelines propose a quantitative methodology for evaluating file formats, identifying those that are in danger of becoming obsolete and choosing those that are more likely to be preserved. To apply this method you need to consider a group of nine factors (Fig. 1) each of which is assigned a numerical value (score) [10], [11], [12], [13]:

a) Standardization (from 0 to 3 points). Formats recognized as de jure standard by a standardization body (such as ISO, UNI, W3C, CEN, ITU, SMTPE, etc.) are awarded with the maximum score, equal to 3; those that have not received this recognition but have become de facto standard, thanks to their widespread diffusion, achieve a score of 2. Finally, a format can also be neither de jure nor de facto standard (score equal to 0), but in this case you should tend to exclude it from the list of acceptable formats.

b) Disclosure (from 0 to 3 points). Open formats, i.e. those whose specifications have been published and made available - possibly also for a fee - obtain the maximum score, equal to 3; on the contrary, closed formats, i.e. those whose specifications have not been made available, achieve the minimum score, equal to 0.

c) Proprietary (from 0 to 4 points). Non-proprietary formats, i.e. those that are not encumbered by intellectual property rights and whose specifications are not managed by a private organization but by a community of developers (for example, the LibreOffice community) or by a standardization body, achieve the maximum score, equal to 4. Proprietary formats, i.e. those that have been created by a private organization (for example, a software house), which owns the intellectual property rights and manages the specifications, achieve a variable score, ranging from 0 to 3. In particular, formats that are proprietary but free to use achieve a score of 3. Proprietary formats who allow the reading of documents already encoded in this format but not the production of new ones achieve a score of 2. Finally, those that do not even allow the reading of documents encoded according to this format achieve a score of 0.

Figure 1 The factors you need to evaluate when calculating the interoperability index.

d) Extensibility (from 0 to 2 points): Extensible file formats, i.e. conceived from the outset to allow subsequent versions that progressively increase their functionality, achieve the maximum score, equal to 2; conversely, those that cannot be extended receive a score of 0.

e) Level of metadata (from 0 to 3 points): it takes into account the level of connection between the document and the metadata associated with it. Formats that allow you to embed metadata within the record get the highest score of 3.

f) Robustness (from 0 to 2 points). Non-robust file formats, i.e. those that, in the case of corruption of the bit stream (the sequence of "0" and "1" of

which an electronic record is made up), do not allow to recover any part of the original file, achieve the minimum score, equal to 0. Completely robust file formats, which include mechanisms to verify any loss of integrity and allow the recovery of the entire document (or, in the case of container formats, of the intact parts), achieve the maximum score, equal to 2. Partially robust file formats achieve a score equal to 1.

g) Device independence (from 0 to 4 points). File formats that are independent from the device, i.e. can be represented in a reliable manner and always in the same way independently from the hardware platform and the software used, achieve the maximum score, equal to 4; those dependent on the device achieve the minimum score, equal to 0.

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

h) Backward and forward compatibility. Backward compatible file formats are those you can use with one of the previous versions of the software that produced them; forward compatible formats are those you can use with any software version subsequent to the one that produced them.

i) Textual or binary encoding. Textual formats are those that allow you to extract the information content by encoding each byte of the binary sequence with the corresponding character (for example, in the ASCII code); conversely, formats that do not allow this type of representation are binary.

Based on this evaluation, you can assign a numerical value to each of the nine factors to evaluate (with the exception of the last two, to which Annex 2 does not assign a specific score).

The Italian Agency for Digital Government called “interoperability index” the sum of these values (although for many this name does not seem very appropriate). It can vary between a minimum of 0 (zero) and a maximum of 21. The most interoperable file format is the one that reaches an index equal to 21; the least interoperable is the one that achieves an index equal to zero. A value equal to 12 is considered as a sufficiency threshold: all those formats that reach an interoperability index equal to or greater than 12 are "acceptable"; lower values show objective problems that must be addressed as soon as possible using, for example, migration processes or other methodologies.

You must pay particular attention when evaluating container formats because in this case you need to assess not only the container format itself but every single digital object contained within it; furthermore, you must consider the lowest (i.e. worst) value for each of these objects. For example, for a multimedia container format, you must evaluate the format of the audio and video files contained in it. For a packet file format you must evaluate all the digital objects included in the package and, if the package includes, in turn, other container formats, you need to evaluate them as well with the same criterion set out above.

III. Some Examples of Calculating the Interoperability Index

To better understand the mechanism for calculating the interoperability index, let us try to apply it to the case of some commonly used file formats. Let's start by considering the DOC format, the default format of Microsoft Word up to version 2003. Table I shows the evaluation of the nine factors you need to consider to calculate the interoperability index.

TABLE I

Calculation of the interoperability index for DOC file format

Factor	Interval	Assessment	Score
Standardization	0 to 3	The DOC format is not a de jure standard but certainly can be still considered a de facto standard	2
Disclosure	0 to 3	The format specifications, initially closed, are now open	3
Proprietary	0 to 4	The format is proprietary	0
Extensibility	0 to 2	The DOC format was abandoned by Microsoft and replaced with the DOCX format	0
Metadata level	0 to 3	Metadata are embedded within it	3
Robustness	0 to 2	Being a binary format, robustness is limited	1
Device independence	0 to 4	Problems with viewing DOC file in environments other than Windows are known	2
Backward and forward compatibility	n. a.	The format is forward compatible	-
Textual or binary encoding	n. a.	The format is binary	-
		TOTAL	11

The score obtained is 11 and it does not reach the sufficiency threshold. Therefore, the format is not interoperable. Note that the evaluation presents a certain degree of subjectivity because different evaluators could assign different scores for each of the factors evaluated.

Let us now consider the DOCX format. Unlike the previous DOC format, which stores the record data in a single binary file, the DOCX format uses the Open Packaging Conventions [14] to create a "package" of files, compressed with the ZIP algorithm, inside which the various components needed to represent the record are collected. Within a DOCX file, there are a [Content-Types].xml file and some folders (such as "docProps", "Word" and "_rels", which contain the properties of the document, the contents and the relationships between files) (Fig. 2).

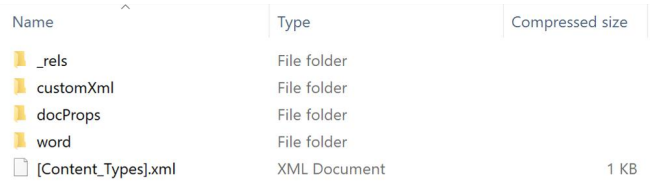

Figure 2 The components within a DOCX file.

This structure makes the document content more accessible. For example, text is saved in a plain text file and images embedded in the document are stored as individual image files. These files may also include page formatting information, author data, and document markups. It is therefore a "package format" and you need to examine the individual components inside the package, which can also be very different. In fact, you can embed not only images (in various formats: JPG, PNG, GIF, etc.) but also audio content (also in various formats: WAV, WMA, MP3, etc.) and video content (also in various formats: WMV, AVI, MP4, etc.). Let us consider, for simplicity, the case of a DOCX file with only textual content. In this case, obviously simplified and only partially representative of the records contained in any digital archive, you can calculate the interoperability index as shown in Table II.

The result obtained is equal to 20, therefore the format, in this simplified case, is interoperable. For completeness, you should repeat the calculation considering the various combinations that can occur in practice. For example, DOCX files can contain also images, audio content and video content, obviously in addition to text. It is possible that unexpected conclusions may be drawn from this analysis: for example, it may happen that the format is interoperable if images are encoded in certain interoperable formats (such as JPG) are incorporated, while it may not be interoperable in the case in which video content in formats that are not interoperable is incorporated.

TABLE II

Calculation of the interoperability index for DOCX file format

Factor	Interval	Assessment	Score
Standardization	0 to 3	The DOCX format is a de jure standard (ISO 29500) and is now also a de facto standard	3
Disclosure	0 to 3	The format specifications are open	3
Proprietary	0 to 4	The format is non-proprietary	4
Extensibility	0 to 2	The format is extensible	2
Metadata level	0 to 3	Metadata are embedded within it	3
Robustness	0 to 2	In this particular case the format contains only textual content, so robustness is high	2
Device independence	0 to 4	DOCX files can be read on various devices, although compatibility issues are sometimes reported	3
Backward and forward compatibility	n. a.	The format is forward compatible	-
Textual or binary encoding	n. a.	In this case the format is textual (encoded in XML)	-
		TOTAL	20

You can make similar considerations for other types of container formats, such as multimedia: in those cases, you need to evaluate all possible combinations with all possible formats of contents you can embed inside the container. Just to give an idea, consider the MKV (Matrioska video) container format, widely used for the creation and storage of audiovisual records. Inside an MKV file you can find video content in various formats (MPEG-2, MPEG-4, WMV, RealVideo, Adobe Flash, H.264, H.265, VP8 VP9, Theora, etc.), audio content also in various formats (PCM, MP2, MP3, AAC, Vorbis, WMA, RA, AC3, DTS, FLAC, etc.), as well as subtitles (USF, SRT, ASS/SSA, OGG WRIT, WebVTT, etc.) and metadata (Fig. 3).

If you want to fully assess the format, you should examine all the possible combinations (remembering that, in the case of the presence of several objects, you need to take into account the worst value) and both the interoperable and non-interoperable ones should be identified. The complexity of the operation lies in the large number of digital objects that can be present inside the container, and, consequently, you need to assess a very large number of possible combinations. Therefore, you can have cases in which, by virtue of a certain combination, the format is interoperable and cases in which, by virtue of different combinations, the format is not interoperable. This example demonstrates the complexity of the matter: you cannot reduce the evaluation of a file format to some simplified cases but you must consider all the possible situations that may occur.

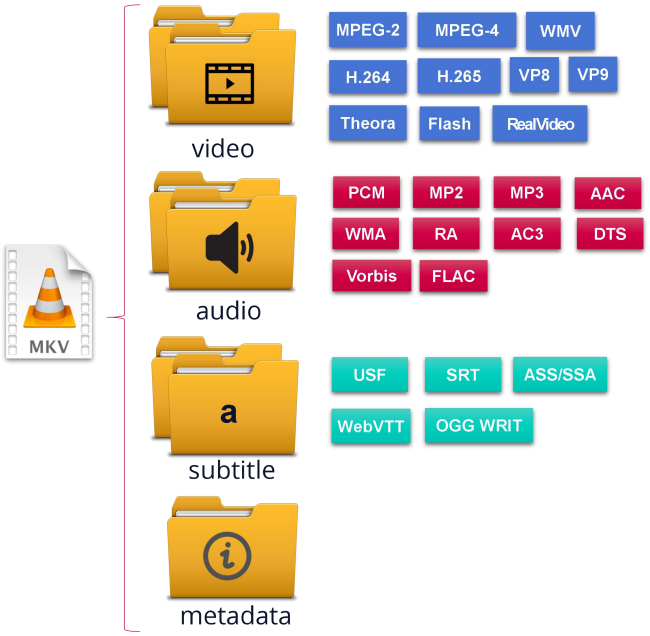

Figure 3 The digital objects that may be present within the MKV format.

Let us now consider the well-known MP3 format, widely used for sound recordings. You can calculate the interoperability index as shown in Table III. Therefore, even if the MP3 format, to date, can be considered interoperable, being a format that is moving towards obsolescence [15], you must kept it 'under control' by repeating the evaluation at least every year.

If, following an interoperability assessment, you find that a file format is not interoperable (and therefore potentially at risk), you need to migrate it to a more interoperable format. In any case, you need to migrate towards file formats that improve interoperability, or at least not make it worse (you can verify this by calculating the interoperability index). In particular, you cannot migrate 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.

TABLE III

Calculating the interoperability index for MP3 file format

Factor	Interval	Assessment	Score
Standardization	0 to 3	The MP3 format is a de jure standard (ISO / IEC 11172-3 and ISO / IEC 13818-3) and also a de facto standard	3
Disclosure	0 to 3	The format specifications are open	3
Proprietary	0 to 4	The format is proprietary but free to use	3
Extensibility	0 to 2	The format is not extensible	0
Metadata level	0 to 3	ID3 tags can be embedded in an MP3 file	3
Robustness	0 to 2	Although MP3 format is based on lossy compression algorithm, it uses internal checksums, so it is partially robust	1
Device independence	0 to 4	MP3 files can be played and any device	4
Backward and forward compatibility	n. a.	The format is forward compatible	-
Textual or binary encoding	n. a.	The format is binary	-
		TOTAL	17

IV. Future Developments

As we have already pointed out, this quantitative methodology for evaluating file formats represents a huge step forward compared to other qualitative approaches. The decision to make the selection of file formats less subjective by evaluating nine file format factors and calculating the numerical value of the interoperability index appears to be correct and valid from a methodological point of view. However, we can give some suggestions, to improve it and make it an even more effective index.

First, some of the factors you need to assess are not easy to evaluate, either because the sources of information are insufficient or do not exist at all. For popular file formats, it is relatively easy to acquire

the information you need for evaluating them, but for the less popular file formats, it may be difficult to find the information you need. Perhaps you need to retrieve and thoroughly analyse the file format specifications that often are highly technical documents and consist of hundreds or thousands of pages. Take, for example, the "extensibility" factor: how can you know if a format was built from the outset to be extensible? In some cases this is not a problem (think of the PDF/A which was created from the outset as extensible format, and this feature has been widely communicated), but in others cases, especially for new file formats, you cannot know in advance if the producer plans to develop further versions of that format. You can often tell if a format is extensible only after some time, when later versions are released. It is also not easy to evaluate the "metadata level" factor. In fact, you have to resort, once again, to the analysis of the format specifications, but it is not certain that the information you need is immediately available since these are often very technical documents. In any case, this kind of search usually requires a lot of time and technical skills that are not within everyone's reach and not infrequently ends with a stalemate because it is not possible to find the desired information.

In some cases, you can assess some factors only experimentally. Let us take, for example, the robustness. You can measure it only carrying out experimental tests, using specific software that artificially "corrupt" the sequence of bits thus simulating the corruption phenomena due to the passage of time, defects in the storage media, unsuccessful transfer operations, errors in the transmission of records through networks, etc. Then you can try to recover the content completely or partially. Again, the ability to conduct such inspections requires skills that, in most cases, far exceed those found among the organization's staff who must carry out the interoperability assessment.

Some factors, such as the "device independence" appear nowadays less useful than in the past, since current file formats are usually device independent. Others factors seem not to be very relevant for interoperability, such as the "extensibility", since the decision to prefer extensible rather than non-extensible formats may not be accepted by everyone.

Finally, some important factors are not included in the list of those you need to evaluate: one of this is the "adoption", that is unanimously considered one of the most important features since the wide adoption of a file format is one of the major deterrents against its obsolescence.

The attribution of the score to the various factors could also be refined, in order to give greater importance to those that actually affect the long-term preservation of file formats and to obtain results closer to reality.

We would like now to return to the issue of subjectivity of evaluations which we mentioned at the beginning and which we consider a particularly critical aspect. In fact, you must consider that the evaluation of the same file format carried out by different organizations could lead to different results, while it would be desirable that the assessment of the same file format would provide the same results regardless of who performs it. This requires that the factors you need to consider are measurable with a low level of uncertainty, but this result is particularly difficult to achieve. Unfortunately, it is not possible to completely eliminate subjectivity in the evaluation process, and this could be a critical problem to solve or, in some way, mitigate.

Some public administrations have suggested that a possible solution to minimize subjectivity could be to establish an "Italian File Formats Registry", managed by a team of experts with specific skills on file formats issues, which will carry out interoperability assessments of file formats and will make them publicly available to all. Some national archives have already implemented similar registries. For example, the National Archives of the United Kingdom have developed the technical registry PRONOM (Public Record Office and NOM) [16], which contains a searchable database of technical information on file formats, together with software tools and the necessary technical environments to access it. In the past, Harvard University Libraries had developed the Global Digital Format Registry (GDFR), a technical registry that ran from 2005 to 2009 [17]. From 2010 to 2016 the Unified Digital Format Registry (UDFR) was also active and tried to "unify" the functions of two existing registries, PRONOM and GDFR. It was the result of a project developed by the University of California Curation Center (UC3) at the California Digital Library (CDL) and funded by the Library of Congress National Digital Infrastructure Preservation Program (NDIIPP) [18]. However, the

currently active registers contain extensive information on file formats, but none of these contains all the information necessary to carry out the interoperability evaluation required by the Italian guidelines.

The Italian registry could initially contain the evaluations of more general use file formats (considering that many of them are common to several organizations) and would gradually be enriched with the evaluations of specific file formats that would be carried out from time to time. The register would be kept constantly updated and could also provide the information about the degree of obsolescence of file formats. For example, it may keep lists of file formats divided into the following categories:

- low risk of obsolescence;
- medium risk of obsolescence;
- high risk of obsolescence;
- obsolete;
- extinct.

Few registries currently make such a distinction, with the exception of a few noteworthy cases, such as the Digital Preservation Coalition that maintains "The 'Bit List' of Digitally Endangered Species" on its website. That list divides the "digital species" into the following categories: Lower risk, Vulnerable, Endangered, Critically endangered, Practically extinct, Concern [19].

V. CONCLUSIONS

In conclusion, the methodology proposed by the Italian Agency for Digital Government for the evaluation of file formats, although it can be improved with some adjustments, appears effective and consistent, and can also be proposed as best practice. Furthermore, although it would certainly make more sense to contribute to one of the existing initiatives, the establishment of an Italian Registry would be very appreciated by the Italian public administrations, since it would be an authoritative and reliable source of information that everyone could draw on. Although in many cases the acceptability of file formats is an institution-specific decision, it would definitely make their assessment easier and more consistent.

REFERENCES

[1] G. Drago, Recommended File Formats for Long-Term Archiving and for Web Dissemination In Phaidra, https://phaidra.cab.unipd.it/static/EN-file-formats.pdf.

[2] Library of Congress, Recommended Formats Statement, https://www.loc.gov/preservation/resources/rfs

[3] Library of Congress, Sustainability of Digital Formats: Planning for Library of Congress Collections, https://www.loc.gov/preservation/digital/formats;

[4] Digital Preservation Coalition, File Formats Assessments, https://wiki.dpconline.org/index.php?Title=File_Formats _Assessments.

[5] The National Archives of United-Kingdom, Selecting File Formats for Long-Term Preservation, https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.

[6] J. Rog, C. Van Wijk, Evaluating File Formats for Long-term Preservation https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.181.506&rep=rep1&type=pdf

[7] United States National Archives and Record Administration (NARA), Digital Preservation Risk Matrix https://github.com/usnationalarchives/digital-preservation/tree/master/Digital_Preservation_Risk_Matrix

[8] Centre de coordination pour l'archivage à long terme de documents électroniques (CECO), Matrice d'évaluation https://kost-ceco.ch/cms/evaluation.html

[9] Italian Agency for Digital Governement, Guidelines on The Creation, Management and Preservation of Digital Records (Linee Guida sulla formazione, gestione e conservazione dei documenti informatici) https://trasparenza.agid.gov.it/archivio19_regolamenti_0_5385.html

[10] DELOS, File Formats Typology and Registries for Digital Preservation, web.archive.org/web/20110721194942/http://www.dpc.delos.info/private/output/DELOS_WP6_d631_finalv2 (5) _urbino. pdf.

[11] InterPARES 2, Selecting Digital File Formats for Long-Term Preservation. General Study 11. Final Report, http://www.interpares.org/ip2/ip2_case_studies.cfm?study=35.

[12] Recommended Preservation Formats for Electronic Records, https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records

[13] Library and Archives Canada, File Format Guidelines for Preservation and Long-term Access, http://www.councilofnsarchives.ca/sites/default/files/LAC%20File%20Format%20Guidelines%20for%20Preservation%20and%20Long-term%20v1_2010-12_0.pdf

[14] Open Packaging Conventions Fundamentals, https://docs.microsoft.com/en-us/previousversions/

windows/desktop/opc/open-packaging-conventions-overview>.

[15] RIP MP3: Another File Format Slips into Obsolescence, https://preservica.com/news/rip-mp3-another-file-format-slips-into-obsolescence

[16] The National Archives of United-Kingdom, The Technical Registry Pronom, https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[17] Stephen Abrams, Establishing a Global Digital Format Registry , Library Trends 54 (1) , June 2005 , DOI: 10.1353/lib.2006.0001 https://www.researchgate.net/publication/32956809_Establishing_a_Global_Digital_Format_Registry

[18] Unified Digital Format Registry (UDFR), https://www.udfr.org

[19] Digital Preservation Coalition, The 'Bit List' of Digitally Endangered Species, www.dpconline.org/digipres/champion-digital-preservation/bit-list.

Note: All URLs were last accessed April 20, 2021.