Review of the NCD 40 (2022), 19–30

Natalie Ostráková, Vojtěch Kopský
Department of Standards, National Library of the Czech Republic

ARCHIVAL FILE FORMAT SELECTION AND DESIGN OF FILE FORMAT SUSTAINABILITY MATRIX FOR THE NATIONAL LIBRARY OF THE CZECH REPUBLIC

Abstract. Selection of sustainable format is essential for long term preservation of digital content. Sustainability factors are known and compliance of formats with those criteria can be assessed with help of abundant resources. There is a considerable consensus on criteria, yet approaches to format evaluation vary amongst institutions. In this article we present discussion of the topic and our approach to such selection. We also propose the evaluation matrix, intended to make the selection process in our library more standardized.

1. Introduction

Poor choice of archival file format with respect to its sustainability can lead to a loss of content or need of its arduous retrieval. Sound choice shall yield a format that would last for a significant period and provide enough time to migrate the content to new format when the necessity comes. Both selection of new format and decision on maintenance of a format that is already used should be based on evaluation of format‘s prospect of rendering significant properties of the original in the distant future.

In the last two decades researchers identified numerous factors influencing that prospect [1, 2], and labeled them sustainability factors. More recently the concept of format evaluation became „textbook material“ and it is included in handbooks like Digital Preservation Handbook [3], which represents comprehensive introduction to digital preservation related aspects of file formats. Our approach was designed with the raster image formats in mind, but most of the concepts apply to all file types.

2. Archival format evaluations in leading institutions

More insight into the problem can be gained by studying materials from institution that are involved in this field most extensively. The Library of Congress (LOC) website contains a review of evaluation factors, and an extensive database of „Format descriptions“ where most of the existing formats are described and examined with respect to each of those factors, which makes it extremely useful resource in case you are faced with unknown format [4]. British Library (BL) „Format Preservation Assessments“ are even more extensive and structured as tagged controlled document, but they were produced for limited number of selected formats only [5], only two of which are raster image formats. Swiss based organization Koordinationsstelle für die Daueshafte Archivierung elektronischer Unterlagen (KOST) provides brief description for selected formats (5 for raster images), including its compliance to sustainability factors [6]. Both BL and KOST give recommendations for preservation actions in format assessments, while LOC outlays its own use, termed „local use“, in its format descriptions, while providing format recommendation separately.

Of the three institutions above, only KOST expresses formats compliance to factors in numeric value in addition to a verbal assessment. Numeric evaluation is also summarized in a matrix that is periodically updated [7]. Two additional organizations, The National Archives and Records Administration (NARA) and Harvard University Library fragmented evaluation factors derived from those coined by LOC into components named indicators.

In the case of NARA the indicators are articulated as questions ascertaining compliance of the format to particular indicators and are attributed numeric value that corresponds to format risk and are displayed in array designated Risk Matrix [8]. Harvard Library expresses the indicators in the form of statement and attributes verbally expressed compliance levels to them [9]. Compliance is also indicated by color (green, yellow, red) in the matrix boxes, which correspond to compliance status (good, neutral, bad).

Last but not least, The Federal Agencies Digitization Initiative (FADGI), the creator of profound Technical Guidelines for Digitizing Cultural Heritage Materials [10] also presents a matrix comparing selected raster formats on its website [11]. There is a considerable emphasize on performance factors in it, while the sustainability factors are addressed briefly and only self-documentation is fragmented to indicators.

Instituce	Library of Congress	Harvard library	NARA	FADGI	KOST-CECO	British Library
Evaluation Factors						
Disclosure	yes	yes	yes	yes	yes	yes
Adoption	yes	yes	yes	yes	yes	yes
Transparency / complexity (1)	yes	yes	yes	yes	videoformats only	yes
Self-documentation	yes	yes	yes	yes	videoformats only	no
External dependencies	yes	yes	yes	no	in preselection (2)	yes
Impact of patents	yes	yes	yes	yes	yes	yes
Technical protection mechanisms	yes	yes	yes	yes	no	yes
Format age	no (4)	yes	yes	no	no	yes
Cost factors	no	yes	no	yes	indirectly (5)	no
Embedded or Attached Content	not stand-alone (6)	no	no	yes	no	yes
Quality and functionality	yes	yes	no	yes	yes	no

Yellow color indicates Sustainability Factors

1) The terms Transparency and Complexity are not synomyms, nor interchangeable, but they do partially overlap as they relate to the same kind of obstacles to work with format. Higher complexity corresponds to lower transparency and vice versa.

2) KOST considers compliance granted in all evaluated formats and does not even discuss it.

3) LOC and Harvard lib. break this factor appart into several partial characteristics (indicators), while KOST observes, whether the format can preserve significant properties of the original or file in previous format.

4) LOC does not treat format age as evaluation factor, but includes it in the format description.
5) KOST does not discuss costs explicitly, but it examines Memory Density which relates to costs.
6) It is not stand alone category. It can be described within the Quality and Functionality category.

Table 1. Comprehensive summary of evaluation factors regarded by aforementioned institutions. Those with yellow highlight are considered Sustainability factors.

3. Evaluation factors – Sustainability factors

Adoption, (Implementation, Distribution). Adoption is seen by all the institutions as essential for the sustainability of the format. If a format is widespread, software will be made and will be accessible to users, community experience will be shared, accumulated and deposited in reports. When there will be further need for tools for creation, validation, migration etc., such demand will be met for widespread formats. While in the description of adoption by LOC existing software support is considered a part of it, BL and KOST regard it as a separate factor, in KOST referred to as Implementation.

Both adoption by memory institutions and adoption by other professionals as well as general public are of course relevant to formats sustainability. Adoption by memory institution is more significant though, because it indicates suitability for the purpose of preservation, i.e. sustainability in that particular environment. NARA even regards autoadoption a standalone indicator.

Disclosure. Availability of the documentation simplifies work with the format and promotes new software to be made. Some of the organizations award maximum points to the formats that are an ISO standard with corresponding documentation.

Legal dependencies (Impact of patents). Legal dependencies can be detrimental to formats prospect for longevity since their assertion prevents (often deliberately) software developers from implementing format access to their products. The best situation, as KOST-CECO claims, is when there is free license for the format which should guarantee, that the legal state of the format will not change unexpectedly.

Self-documentation. Capability of the format to carry metadata necessary for its proper rendering, identification and documentation of its origin and purpose is appreciated by LOC as it aids to managing the files. NARA indicators for this factor ask specifically for capacity to embed al types of metadata with compliance to international standards.

External dependencies. While evaluating this factor, LOC and NARA ask whether the format is dependent on specific playback or rendering hardware, software environment, like operating system, plug-ins, scripts or proprietary software. On the other hand in British library terminology external dependencies refer to external content, like external fonts etc. All of those are of course disadvantageous.

Transparency / complexity. Transparency refers to accessibility of the format by common tools and human readability. NARA upholds the broadest scope of the factor, demanding standard character and other encoding, availability of software usable across the hardware and operating system platforms, as well as documentation relevant to format identification and validation, which obviously constitutes overlaps with software support (i.e. adoption), disclosure and external dependencies factors. On the other hand BL concerns itself only with complexity.

With regard to complexity most institutions emphasize compression as a major contributing factor. File formats which do not use compression of the data are seen as less complex i.e. more transparent and if there is a compression, then it depends how well the compression algorithm is understood.

Technical Protection Mechanisms. Presence of encryption and other DRM (digital rights management) mechanisms designed to prevent copying of the file is of course impediment to preservation purposes since it threatens rendering.

Format Age (Lifetime). While Harvard Library considers format age a positive indicator within Adoption Factor, NARA sets it as a separate factor, split into two indicators, time since creation and time since last update and considers both a risk i.e. older the format is, worse the risk. BL addresses the format age in more broadly outlined assessment of Development status.

Embedded or Attached Content. Possibility of embedding or attaching content to file in particular format is listed by BL as a sustainability threatening factor. There are no issues for this factor for JPEG2000, yet for TIFF the BL warns against deviation from Baseline TIFF 6. LOC does not list it as a standalone factor, and discusses it within Functionality factors.

Cost factors (Financial concerns). KOST considers the financial concerns as a part of reason for supporting compression in formats which it rewards in storage density factor. Harvard University addresses tool costs in addition to storage costs. FADGI breaks the cost factors into several separate indicators.

4. Evaluation factors - Quality and functionality factors

Rendering the significant properties of the original is as important characteristic of a format as survival through time, therefore not just sustainability factors are considered when institutions are deciding on preferred format. Those factors differ in between format types, according to type of content. For raster image formats those would be mostly resolution and color management. Since this paper is focused on sustainability, we are not going to discuss other factors any further. Many institutions including ours address those factors in preselection and only the formats that fully comply with technical parameters demand are assessed for sustainability.

5. Previous approaches to format selection in National Library

The tool or method for file format evaluation has not been created in National Library before. Until recently, the standard procedure for choosing the format for archiving purposes consisted of several steps described in following sections.

a. Study of file format recommendations. The first step in selection of the format suitable for our long term preservation purposes was study of file format recommendations published by recognized authorities.

The LOC was already mentioned in connection with their database of format descriptions. It is also issuing Recommended Formats Statement [12] annually since 2014. It contains archival file format recommendations for digital as well as analog formats and updates are accompanied with a „Change history“ document.

Recommendations and guidelines from Federal Agencies Digital Guidelines Initiative (FADGI), is extensive document covering all the technical details of digital preservation [9]. While LOC distinguishes preferred and accepted formats, analogically to format policies we will demonstrate later, FADGI only recommends the optimal formats while distinguishing one to four stars quality levels, consisting of resolution, bit depth, color space etc.

Additional two recommendations we relied on are those of BL and KOST that we have already mentioned since those are included in their format assessments.

b. Study of format policies. Based on the study of the sources mentioned before, smaller group of candidate file formats was obtained, that seemed to be suitable for archival purposes. These file formats were subsequently monitored in format policies of memory institutions. This step enabled us to ascertain the level of adoption of the format.

Format policies are expressions of preferences of individual libraries in their digital preservation. Its individuality is clearly manifested by differences in terms for categories of acceptance.

raster formats considered for archival use >		TIFF (unc.)		TIFF (cmp.)		JP2 (l.less)		JP2(lossy)		GIF		PNG		JPG		BMP		categories of acceptance	
Institution	released	HI	ME	HI	ME	HI	ME	HI	ME	HI	ME	HI	ME	HI	ME	HI	ME	HIGH	MEDIUM
Alabama uni. lib.	2016	Sup					Kno		Kno	Sup		Sup		Sup				Supported	Known
Arthur Lakes Library	2018	Full					Lim		Lim		Lim		Lim	Full			Lim	Full support	Limited supp
Boston Uni Libraries	2011	For					Bit			For			Bit	For			Bit	Format	Bit-level
Canada library and archives	2015	Pre				Pre					Acc	Pre			Acc			Preferred	Accepted
Connecticut uni. lib.	2018	Sup								Sup		Sup		Sup			Kno	Supported	Known
Cornell University Li	2019	Hi			Me	Hi			Me		Me	Hi			Me		Me	High	Medium
Deep Blue (Michigan uni)	2011	L1					L2						L2	L1				Level 1	Level 2
Florida uni. libraries	2012	Hi			Me	Hi			Me		Me	Hi			me		Me	High	Medium
Hawai'i uni.	2019	Sup								Sup		Sup		Sup			Kno	Supported	Known
Houston uni, TX	2018	Hi					Me				Me		Me		ME			High	Medium
LOC	2016	Pre				Pre		Pre		Pre		Pre		Pre		Pre		Preferred	Acceptable
Minnesota University	2014	Full					Lim				Lim		Lim	Full			Lim	Full support	Limited supp
National Archives (USA)	2019	Pre				Pre					Acc	Pre			Acc			Preferred	Acceptable
North Carolina State Archives	2012	Rec				Rec					Acc		Acc		Acc			Recommended	Acceptable
North Carolina State University	2018	Sup								Sup		Sup		Sup			Par	Supported	Partially suppo

Libraries																			rted
Northwestern University	-	Hi.r.			Mo.r.	Hi.r.			Mo.r.				Mo.r.		Mo			Highly recom.	Moderately recom.
Purdue University Libraries	2012	Sus				Sus					Sup		Sup		Sup			Sustainable	Supported
McMaster, Ontario, CAN	-	Full											Bit		Bit			Full	Bit level preserv.
Simon Fraser University Archives	2017	Pre				Pre				Pre		Pre		Pre			Acc	Preferred	Acceptable
Smithsonian archives	-	Pre					Acc						Acc		Acc			Preferred	Acceptable
Southern Illinois uni.	2008	Sup								Sup		Sup		Sup			Kno	Supported	Known
Tasmanian Archives	2015	Rec				Rec					Acc	Rec			Acc			Recommended	Acceptable
Texas A&M uni., TX	2014	Pre					Acc								Acc			Preferable	Acceptable
UK Data service	2014	Rec			Acc		Acc				Acc		Acc		Acc		Acc	Recommended	Acceptable
W - University of Washington Libraries	2014	HI			ME	HI			ME		ME	HI			ME		ME	Highest	Medium
		24			5	11	9	1	6	8	12	13	10	11	13	1	12		

abbreviations: unc.=uncompressed, cmp.= compressed, l.less= lossless

Table 2. Format policies of institutions that display them on their websites

c. Testing of software tools. The primary purpose of software testing in our library is certainly not the format evaluation. We need the software to maintain the files in our workflow and we also advice smaller libraries on freeware suitable for their needs. The search for it provides additional information on adoption though. We seek, test and use software for:

- Creation (codecs, editing sw.)
- Display, rendering, playback, opening
- Migration to another format
- Validation
- Identification
- Metadata extraction
- For the archival file format, all of those programs must be available and functional.

6. New approach - design of the Format evaluation matrix for NL CZ

Despite the original process yielding quite suitable file formats, we felt the need for further standardization of the process that would allow us to fit the evaluation to our specific needs as some of the features followed by other institutions do not concern us. For instance, we chose the sidecar approach to descriptive metadata, where metadata are stored in XML files compliant to metadata schemes placed in dedicated folder in our AIP (Archival Information Package), therefore the possibility of embedding them in the file is not crucial to us. For financial reasons we accept compression, yet we insist on lossless. Also, we address performance and quality factors in preselection of formats, since compliance to those is essential. Therefore, in our matrix, we evaluate sustainability factorsonly.

Standardization will also allow us to follow development of the format status in time, to compare status of competing formats and for reproducibility of the evaluations and transferability of the method to new colleagues. Last but not least, evaluation matrix would be useful for assessment of unknown formats submitted ad hoc.

For each factor we have several indicators that correspond to concrete conditions the format must/should conform to. Each indicator can be formulated as a question with yes or no answer and yes is awarded points corresponding to indicators weight.

7. Results of application to raster formats

In the case of raster formats, the matrix did not serve us as a tool for format selection. We chose JPEG2000 (JP2) as an archival master format based on its performance and acceptance in preservation community long before we designed the matrix. It was used to check the format on the sustainability prospect, and the follow up is intended. Table 3.shows comparison of JP2and TIFF, which is the most preferred format in archiving altogether. Some shortcomings of JP2 can be noticed, which we consider a trade of for the lossless compression.

The third format evaluated was a proprietary RAW, that our vendor suggested we should use for archiving. The general term is used intentionally because all the proprietary RAWs lack features required for archiving, namely disclosure, adoption and transparency, therefore the results would be the same for all of them.

Factor	Indicator	Rationale		JPEG 2000	TIFF uncompr.	propriet. RAW
Disclosure			Weight			
	Format specification is available	The parameters of the format are described and the future user (software creator, etc.) is able to find that information.	1	1	1	0
	Specification is complete and comprehensible	The specification is sufficient source for understanding the format. Future developers are able to create suited software based on it.	1	1	1	0
	Specification is standardized	Specification has undergone revision, external assesment by reliable standardization institution and is preserved. Probability of future accesibility is	1	0,5	0,5	0

		increased.				
	Format is described in format registry of LOC	Format was assesed by an institution experienced in format evaluation. Informations about the format status are recorded including links to additional information sources.	1	0,5	0,5	0
Adoption						
	Format is widely adopted among users	If the format is widespread, there will be interest in development of tools for that format in the future. There will be experts on the format among the stakeholders. Format status should not change sudenly and unxepectedly, and if it does, there shall be a community counteraction. In present we can expect sufficient amount of tools for the format.	1	1	1	0
	Format is used for archiving purposes	Format was chosen by competent institution for archiving purposes i.e. its suitability was already assesed and the choice of format was based on it. There is experience with format use for archiving and there will be interest in development of migration tools for it.	1	1	1	0
	Format is listed in recognized recomendation	Format has undergone thorough evaluation by a recognized subject in digital archiving and therefore there is higher chance it is really suitable for achiving.	1	1	1	0
	Format is listed in format policies of multiple institutions	Format was subject to local evaluations, institutions archive it and accept it with that intention. Institutions examine the risk of format regularly and will be adressing its migration in the future.	1	0,5	0,5	0
Implementation						
	There are several tools for creating files in this format	Abundance of tools indicates that development of such software is feasible. At	1	1	1	0

		the same time, the users are not dependent on one tool and therefore one subject that maintains it.				
	Source code of the file creating tool is available as open source	Using the tool can be relatively inexpensive. One open source implementation can facilitate development of other implementations.	1	1	1	0
	There are multiple rendering tools	The format allows for regular work. Relevant tools can be made easilly and users are not dependent on one tool and institution that maintains it.	1	1	1	1
	Source code of rendering tool is available as open source	Access to content is afordable. One open source implementation can stimulate creation of further ones.	1	0,5	0,5	0

Factor	Indicator	Rationale		JPEG 2000	TIFF uncompr.	propriet. RAW
Implementation			Weight			
	Tool for validation is available	Validation (i.e. revision, whether the format is compliant to its specification) is considered important step in long term preservation. Future software developers can rely on specification when creating aplications, so it always would be possible to use valid files.	1	1	1	0
	Tool for migration to new format is available	In case of risk, the format can be migrated instantly.	1	1	1	1
Transparency/ complexity						
	Format can be identified reliably	A correctly identified format can be continuously monitored, evaluated, extracted from the repository and required operations can be performed with it.	1	1	1	0

	Format can be validated reliably	It can be verified whether the format complies to its specification and therefore it can be maintained with less concern about unexpected complications. The weight is reduced because validation tools are constantly improved and new ones are beeing developed and at the time of evaluation the tool can be still in development. This must be monitored, but the format can still be suitable for archiving purposes.	1	0,5	0,5	0
	Compression is not used	Format is easily readable.	1	0	1	0
	Compression algorithm is known and widespread	Format can be understood.	1	0,5	0	0
Legal dependencies						
	License is royalty free	More software is available and it should be cheaper.	1	0,5	0,5	0
	Format is open source	Use is costless. Promotes format spreading.	1	0	0,5	0
External dependencies						
	Format can be used on multiple operation systems	Format can be implemented in various workflows. Users can work with the format, resp. render it, without switching OS. Format can be expected to work on both contemporary and future platforms.	1	0,5	0,5	0,5
	The format is independent of specific HW	There is no need to purchase and preserve specific HW.	1	1	1	1
	Format can be used in multiple SW aplications	The use of the format does not depend on a single software and is not threathened by its disconection.	1	1	1	1
	Format does not require plug-ins, scripts or external content	State of plugins and externtal content tends to be uncertain, it can vanish overnight. Need for plugins for browsers means that the format is not directly supported by the environment, and there is less developer interest in it.	1	0	0,5	0

abbreviations: uncompr.= uncompressed, propriet.= proprietary

Table 3. National Library Evaluation Matrix Raster Format Comparison

8. Discussion

We are aware that JP2 does not comply with as many indicators, as the TIFF does, but fortunately, all of those are minor ones. JP2uses compression which makes it less transparent, but it's a price we pay for saving the data space. JP2 is not fully open source and you have to use a commercial software if you want a good one. Also some browsers require a plug in to render JP2, which is considered external dependency, and also a marker of lesser adoption. On the other hand it does not have direct impact on the use of the format as archival master. Nevertheless, if the future checkups reveal more serious deficiencies, migration to new format should be considered.

9. Conclusion

We intend to perform periodical evaluations with the use of our matrix on the formats we already use, to provide us with early warning, in case the format status would change. We will also evaluate formats with promising potential that might eventually replace the ones that became obsolete. Finally, the matrix may serve for ad hoc evaluations of formats that would be suggested by external stakeholders.

Shortly after we have finished our first version of evaluation matrix, our department was tasked with evaluation of the proprietary RAW mentioned above. It was of course unsuitable for long term preservation at the first glance, but running it through the matrix took us less than half a day and provided us with much more authoritative position in rejecting the format.

After we gain sufficient experience with using the matrix, it will be undoubtedly subject to revisions. We also hope this exploration might elucidate some new relations between the format properties and its longevity.

References

1. Rog, Judith and Caroline van Wijk. 2008. Evaluating File Formats for Long-term Preservation [online]. 2008 [Accessed: 2022-03-15]. Available at: https://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf
2. Brown, Adrian 2008. Selecting file formats for long-term preservation The National Archives (UK) Digital preservation guidance note 1. [Accessed: 2022-03-15]. Available at: http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf
3. Digital Preservation Coalition. File formats and standards. Digital Preservation Handbook [online]. 2nd ed. Glasgow: Digital Preservation Coalition, 2015 [Accessed: 2022-03-16]. Available at: https://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards
4. Library of Congress, Sustainability Factors. Sustainability of Digital Formats: Planning for Library of Congress Collections [online]. Washington (DC): The Library of Congress, Last Updated: 01/05/2017 [Accessed: 2022-03-15]. Available at: https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml
5. British Library. File Formats Assessments. DPC [online]. 2021 [Accessed: 2022-03-15]. Available at: https://wiki.dpconline.org/index.php?title=File_Formats_Assessments

6. Koordinationsstelle Für Die Dauerhafte Archivierung Elektronischer Unterlagen, 2019 Kriterienkatalog [online]. KOST-CECO, Version 6.0, Juli 2019 [Accessed: 2022-03-15]. Available at: https://kost-ceco.ch/cms/Kriterienkatalog.html
7. Koordinationsstelle Für Die Dauerhafte Archivierung Elektronischer Unterlagen, 2021 Evaluation matrix. KOST-CECO, Version 6.2, December 2021 [Accessed: 2022-06-22]. Available at: https://kost-ceco.ch/cms/Bewertung.html
8. The National Archives and Records Administration, File format Risk Matrix 2021 [Accessed: 2022-03-15]. Available at: https://github.com/usnationalarchives/digital-preservation/blob/master/Digital_Preservation_Risk_Matrix/NARA_File_Format_Risk_Matrix_20211223.xlsx
9. Goethals, Andrea, 2016b. Format matrix tool. Harvard Wiki [online]. Harvard College, 2016 [Accessed: 2022-03-17]. Available at: https://docs.google.com/spreadsheets/d/1buM2XZtkc09kUtUo0W5s0lt4lK_6LALF6VooCJDdQZ0/edit
10. Federal Agencies Digitization Initiative. Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files [online]. Washington (DC): FADGI, September 2016 [Accessed: 2022-03-17]. Available at: http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf
11. Federal Agencies Digitization Initiative. Raster Still Images for Digitization: A Comparison of File Formats [online]. Washington (DC): FADGI, 2014 [Accessed: 2022-03-16]. Available at: https://www.digitizationguidelines.gov/guidelines/raster_stillImage_compare.html
12. Library of Congress. Recommended Formats Statement 2020–2021 [online]. Washington (DC): The Library of Congress, [2020] [Accessed: 2022-03-15]. Available at: https://www.loc.gov/preservation/resources/rfs/RFS%202020-2021.pdf

Vojtech.Kopsky@nkp.cz