Evaluating File Formats for Long-term Preservation

Judith Rog, Caroline van Wijk; National Library of the Netherlands; The Hague, The Netherlands;
judith.rog@kb.nl, caroline.vanwijk@kb.nl

Abstract

The KB has developed a quantifiable file format risk assessment method. This method can be used to define digital preservation strategies for specific file formats. The method contains seven sustainability criteria for file formats that are weighed for importance. There seems to be consensus on the sustainability criteria. However, as the weighing of these criteria is connected to an institution's policy, the KB wonders whether agreement on the relative importance of the criteria can be reached at all. With this paper, the KB hopes to inspire other cultural heritage institutions to define their own quantifiable file format evaluation method.

Introduction

Over more than a decade, the Koninklijke Bibliotheek (KB) has been involved with the preservation of digital publications. In 1996, the first agreements were signed with Elsevier Science and Kluwer Academic, international publishers of Dutch origin, on the long-term preservation of their e-journals. In 2002 it was decided that the scope of the e-Depot would be broadened to cover the whole spectrum of international scientific publishing. The e-Depot, the electronic archive the KB uses for the long-term storage and preservation of these journals, became operational in 2003 [1]. At this moment, the e-Depot holds over 10 million international e-publications. Up until now, the vast majority of the publications in the e-Depot consist of articles from e-journals. For all but a few of these articles the format in which they are published is the Portable Document Format (PDF), ranging from PDF version 1.0 to 1.6. For this reason, the research the KB has done to keep the articles preserved and accessible for future use, focused mainly on PDF. At this moment, however, the scope of the e-Depot is broadened. Apart from the ongoing ingestion of the electronic publications, in the coming five years, data resulting from ongoing projects such as web archiving [2], DARE [3], national e-depot [4] and several digitisation projects [5] will be ingested in the e-Depot as well. The content from these projects such as DARE, web archiving and national e-depot is very heterogeneous concerning file formats. Even the 'traditional' publications that the publishers are providing are getting more and more diverse. Articles can be accompanied by multi media files or databases that illustrate the research.

This more diverse content forces the KB to reconsider its digital preservation strategy. At the foundation of each strategy is the basic principle that the KB will always keep the original publication. The digital preservation strategy describes what actions (e.g. migration or emulation) the KB undertakes to ensure that these publications are preserved and remain accessible for future use. The strategy also describes which choices to make for specific formats during creation, ingest or at a later stage because choices at each of these stages can influence the sustainability of the file. The current strategy is mainly focused on preserving PDF files, but our strategy will need to cover a much wider variety of formats from now on. Whether preservation actions are needed, and which actions are needed depends among other things on the long-term sustainability of the file format of the publication. But what makes a file format suitable for long-term preservation? The criteria for evaluating file formats have been described by several authors [6], [7], [8], [9], [10]. But only very rarely though are these

criteria applied to a practical assessment of the file formats [11]. To apply the sustainability criteria we need to know whether all criteria are equally important or whether some are more important than others. And how do you measure whether, and to what degree the format meets the criteria? The application of the criteria should be quantifiable to be able to compare file formats and to give more insight into the preference for certain file formats for long-term preservation.

The KB has started to develop such a quantifiable file format risk assessment. The file format risk assessment facilitates choosing file formats that are suitable for long-term preservation. This paper describes the file format assessment method that the KB has developed and how it is applied in the preservation strategies at the KB. The KB invites the digital preservation community to start a discussion on sustainability criteria and the importance of each criteria by presenting its file format evaluation method.

File Format Assessment for Long-term Preservation

Methodology

The general preservation criteria used in the KB's method originate from the aforementioned digital preservation literature. The KB's assessment method does not take into account quality and functionality criteria such as clarity or functionality beyond normal rendering as defined in Arms & Fleischhauer [9]. The KB archives publications which are end products that for example do not need editing functionality after publishing. Also the KB archives the publications for long-term preservation purposes and is not the main point of distribution for these publications. Regular access to and distribution of publications is offered by publisher's websites and university repositories etc. This reasoning might be very specific to the KB and it explains the choice for only applying sustainability criteria in the risk assessment method. In the next sections, the criteria, the weighing of the criteria and an example of the application of the method will be described.

The criteria on which classifications of suitability of file formats from the view point of digital preservation will be based are described below. The criteria form measurable standards by which the suitability of file formats can be assigned. The criteria are broken down into several characteristics that can be applied to all file formats. Values are assigned to each characteristic. The values that are given differ among file formats. The sustainability criteria and characteristics will be weighed, as the KB does not attribute the same importance for digital preservation planning to all characteristics. The weights that are assigned to the criteria and their characteristics are not fixed. They depend on the local policy of an institution. The weights that are used in the examples in this paper are the weights as assigned by the KB based on its local policy, general digital preservation literature and common sense. The range of values that can be assigned to the characteristics are fixed.

The weighing scale runs from one to seven. These extremes are arbitrary. Seven is the weight that is assigned to very important criteria from the point of view of digital preservation and zero is the score assigned to the least important criterion. The values that are assigned to the characteristics range from zero to two. The lowest numerical value is assigned to the characteristic value that is seen as most threatening to digital preservation and long-term accessibility. This value is zero. The highest numerical value is assigned to the characteristic value that is most important for digital preservation and long-term accessibility. This value is two. The scale from zero to two is arbitrary. The criteria do not all have the same number of

characteristics. The total score that is assigned to all characteristics is therefore normalised by dividing the score by the number of characteristics.

By applying the file format assessment method to a file format, the format receives a score that reflects its suitability for long-term preservation on a scale from zero to hundred. The higher the score, the more suitable the format is for long-term preservation. The score a format receives can vary over time. A criterion such as *Adoption* for example is very likely to change over time as a format gets more popular or becomes obsolete.

Criteria defined

The criteria that are used in this methodology are *Openness, Adoption, Complexity, Technical Protection Mechanism (DRM), Self-documentation, Robustness and Dependencies.*

Openness

The criterion *Openness* of a file format is broken down into the characteristics *Standardisation, Patents, Reader with freely available source*. These characteristics indicate the relative ease of accumulating knowledge about the file format structure. Knowledge about a file format will enhance the chance of successful digital preservation planning.

Adoption

The criterion *Adoption* of a file format has only one characteristic: *Market share*. This characteristic indicates the popularity and ubiquity of a file format. When a specific file format is used by a critical mass, software developers (commercial, non commercial) have an incentive to sustain support for a file format by developing software for the specific file format such as readers and writers.

Complexity

The characteristic *Complexity* of a file format is broken down into the characteristics *Human readability*, *Compression*, *Versatility*. These characteristics indicate how complicated a file format can be to decipher. If a lot of effort has to be put into deciphering a format, and with the chance it will not completely be understood, the format can represent a danger to digital preservation and long-term accessibility.

Technical Protection Mechanism (DRM)

The characteristic *Technical Protection Mechanism* of a file format is broken down into the characteristics *Password protection, Copy protection, Digital signature, Printing protection* and *Content extraction protection*. These characteristics indicate the possibilities in a file format to restrict access (in a broad sense) to content. Restricted access to content could be a problem when the digital preservation strategy migration is necessary to provide permanent access to the digital object.

Self-documentation

The characteristic *Self-documentation* of a file format is broken down into the characteristics *Metadata* and *Self describing format*. These characteristics indicate the format possibilities concerning encapsulation of metadata. This metadata can be object specific or format specific. When a format facilitates the encapsulation of object specific information (such as author, description etc.) or format specific information in the header on how to read the format for example, the format supports the preservation of information without

references to other sources. The more that is known about a digital object, the better it can be understood in the future.

Robustness

The characteristic *Robustness* of a file format is broken down into the characteristics *Robust against single point of failure, File corruption detection, File corruption correction, File format stability, Backward compatibility, Forward compatibility* and *Format can only be changed via a defined procedure*. These characteristics indicate the extend to which the format changes over time and the extend to which successive generations differ from each other. Also, this characteristic provides information on the ways the file format is protected against file corruption. A frequently changing format could threaten continuity in accessibility for the long term. Large differences among generations of a file format could endanger this continuity equally. The values for file format stability are rare release of newer versions, limited release of newer versions and frequent release of newer versions correspond to release once in ten years, release once in five years and release once a year respectively.

Dependencies

The characteristic *Dependencies* of a file format is broken down into the characteristics *Not dependent on specific hardware, Not dependent on specific operating systems, Not dependent on specific reader* and *Not dependent on other external resources*. These characteristics indicate the dependency on a specific environment or other resources such as fonts and codecs. A high dependency on a specific environment or on external resources provides a risk for digital preservation and long-term accessibility. External resources could be lost over time and difficult to retain and a high dependency on a specific environment strongly ties the format to a specific time and space.

The full list of criteria, the weights as assigned by the KB, the criteria and their possible values can be found in Appendix I. An example of the file format assessment method applied to PDF 1.4 and PDF/A can be found in Appendix II

Application of File Format Assessments

The KB has defined a digital preservation policy for the content of the e-Depot. This policy is the starting point for digital preservation strategies for the digital objects stored in the e-Depot. A digital preservation strategy starts at creation time of a digital object and defines preservation actions on the object at a later stage in the object's life cycle. The KB will not restrict the use of specific file formats for deposit. Any format in general use can be offered. However, KB does give out recommendations and uses the file format assessment method to define strategies.

During the last decade the KB has carried out many digitisation projects. The development of digitisation guidelines has been part of these projects. These guidelines not only make sure that specific image quality requirements are met. They also ensure that the created master files meet the requirements that the digital preservation department has set for metadata and technical matters such as the use of specific file formats and the use of compression (no compression or lossless compression). A file format evaluation method is essential for making well thought-out choices for specific file formats at creation time of digital objects.

The KB has had a lot of influence on the creation process as the owner of the digitisation master files. However, this is not the case for millions of digital publications that have been and will be deposited by international publishers. The KB does have deposit contracts that contain several technical agreements (e.g. file format in which the publisher chooses to submit the publications). Also, as most publications are deposited in the PDF, guidelines for the creation of publications in PDF [12] have been created. The PDF guidelines are related to the standard archiving format PDF/A, but are easier to read for non-technical persons. They contain ten 'rules' for PDF functionality that describe best practices at creation.

As was mentioned before, the deposited publications have been quite homogenous concerning file format. Most publications have been deposited in the PDF version 1.0 to 1.6. The file format assessment method has been used to assess this main format stored for its digital preservation suitability. However, new projects will make the digital content of the archive more heterogeneous in the near future. This will require more elaborated file format evaluations.

One example of the use of file format evaluations for new e-Depot content is the evaluation of formats that are harvested for the DARE project. DARE publications are harvested from scientific repositories such as the Dutch university repositories. Most harvested publications are PDFs, however a small part of the articles are harvested in MS Office document formats such as MS Word and MS PowerPoint and in the WordPerfect format. The concrete result of the use of file format risk assessment at the KB is the decision to normalise MS Office documents and WordPerfect documents to a standard archiving format: PDF/A. MS Word documents score 44% if assessed by the assessment method. PDF/A's assessment score amounts to 89 %. The main difference between the formats can be found in the criteria *Openness* and *Dependencies*. For these two criteria, MS Word does have a considerably lower score than PDF/A has. In accordance with the preservation policy both original and normalised files are stored for long-term preservation purposes.

Interestingly enough, an archival institution that is partner in the National Digital Preservation Coalition (NCDD), does not consider PDF/A suitable for archiving its digital data for the long term. One of its valid arguments for not using PDF/A was that PDF/A does not offer the same editing functionality that is available in datasheets. It would be very interesting to compare the differences among cultural heritage institutions concerning the sustainability criteria and the importance of these criteria. This will be much easier if institutions make their file format evaluation quantifiable.

The biggest challenge for the application of the file format risk assessment in the near future will be the web archiving project. As websites contain many different file formats, this new type of content for the e-Depot will require quite different preservation strategies and plans from the current ones.

Conclusion and Discussion

This paper describes the file format assessment that was developed by the KB to assess the suitability of file formats for long-term preservation. The suitability is made quantifiable and results in a score on a scale from zero to hundred that reflects the suitability of the format for long-term preservation. Formats can easily be compared to each other. The criteria, characteristics and scores that the formats receive are transparent.

The KB hopes to receive feedback on the methodology from other institutions that have to differentiate between formats to decide which format is most suitable for long-term

preservation. There seems to be consensus on the sustainability criteria. However, the KB would like to know whether these criteria are the right ones and whether the possible scores a format can receive on a characteristic offer practical options to choose from. The weighing that can be applied to a criterion is not fixed in the methodology. The weighing can be adjusted to the local policy. Therefore the KB would like to invite other cultural heritage institutions for a discussion about and preferably a comparison of quantifiable file format risk assessments.

References

[1] "The archiving system for electronic publications: The e-Depot", National Library of the Netherlands, Information available at: http://www.kb.nl/dnp/e-depot/dm/dm-en.html;
[2] "Web archiving", Digital Preservation Department National Library of the Netherlands. Information available at: http://www.kb.nl/hrd/dd/dd_projecten/projecten_webarchiveringen.html;
[3] "DARE: Digital Academic Repositories", Digital Preservation Department National Library of the Netherlands. Information available at: http://www.kb.nl/hrd/dd/dd_projecten/projecten_dare-en.html;
[4] "Online deposit of electronic publications", National Library of the Netherlands. Available at: http://www.kb.nl/dnp/e-depot/loket/index-en.html;
[5] "Digitisation programmes & projects", National Library of the Netherlands. Information available at: http://www.kb.nl/hrd/digi/digdoc-en.html;
[6] Folk, M. & Barkstrom, B. R., "Attributes of File Formats for Long-Term Preservation of Scientific and Engineering Data in Digital Libraries", NCSA, 2002. Available at: http://www.ncsa.uiuc.edu/NARA/Sci_Formats_and_Archiving.doc;
[7] Christensen, S.S., "Archival Data Format Requirements", Netarchive.dk, 2004, Available at: http://netarchive.dk/publikationer/Archival_format_requirements-2004.pdf;
[8] Brown, A., "Digital Preservation Guidance Note 1: Selecting File Formats for Long-Term Preservation", National archives of the United Kingdom, 2003, Available at: http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf;
[9] Arms, C. & Fleischhauer, C., "Digital formats: Factors for sustainability, functionality and quality", presentation IS&T Archiving, 2005;
[10] "Sustainability of Digital Formats Planning for Library of Congress Collections", Library of Congress, Available at: http://www.digitalpreservation.gov/formats/sustain/sustain.shtml;
[11] Anderson, R., Frost, H., Hoebelheinrich, N. & Johnson K., "The AIHT at Stanford University", D-Lib Magazine December 2005. Available at http://www.dlib.org/dlib/december05/johnson/12johnson.html;
[12] Rog, J., "PDF Guidelines: Recommendations for the creation of PDF files for long-term preservation and access", Available at: http://www.kb.nl/hrd/dd/dd_links_en_publicaties/PDF_Guidelines.pdf.

Author Biography

Caroline van Wijk (1973) has a BA degree in Arts from the Rietveld Academy. She finished a Java software engineer training in 2000. Directly after, she had been working at a number of web development companies for well over four years before she joined the KB in 2004. At the KB, she had worked on the pilot project Tiff-archive as the technical project staff

member until December 2005. Since 2006, she leads the migration implementation project and takes part in the European project Planets as a digital preservation researcher and work package leader.

Judith Rog (1976) completed her MA in Phonetics/Speech Technology in 1999. After working on language technology at a Dutch Dictionary Publisher she was employed at the National Library of the Netherlands/Koninklijke Bibliotheek (KB) in 2001. She first worked in the IT department of the KB for four years before joining the Digital Preservation Department in 2005. Within the Digital Preservation Department she participates in several projects in which her main focus is on file format research.

Appendix I

Table 1: All criteria, weighting factors, characteristics and values that can be applied

Criterion (weighing factor)	Characteristic	values	
Openness (7)			
	Standardisation		
		2	De jure standard
		1,5	De facto standard, specifications made available by independent organisation
		1	De facto standard, specifications made available by manufacturer only
		0,5	De facto standard, closed specifications
		0	No standard
	Patents		
		2	No patents
		1	Partially patented
		0	Heavily patented
	Reader with freely available source		
		2	Freely available open source reader
		1	Freely available reader, but not open source
		0	No freely available reader
Adoption (6)			
	Market share		
		2	Widely used
		1	Used on a small scale
		0	Rarely used
Complexity (3)			
	Human readability		
		2	Structure and content readable
		1	Structure readable
		0	Not readable
	Compression		
		2	No compression
		1	Optional compression
		0	Always compressed
	Versatility		
		2	Not versatile
		1	Some versatility
		0	Very versatile
Technical Protection Mechanism (DRM) (3)			
	Password protection		
		2	Not possible
		1	Optional
		0	Mandatory
	Copy protection		
		2	Not possible
		1	Optional
		0	Mandatory
	Digital signature		

Criterion (weighing factor)	Characteristic	values	
		2	Not possible
		1	Optional
		0	Mandatory
	Printing protection		
		2	Not possible
		1	Optional
		0	Mandatory
	Content extraction protection		
		2	Not possible
		1	Optional
		0	Mandatory
Self-documentation (1)			
	Metadata		
		2	Possibility to encapsulate user-defined metadata
		1	Possibility to encapsulate a limited set of metadata
		0	No metadata encapsulation
	Self-describing format		
		2	Fully self-describing
		1	Partially self-describing
		0	Not self-describing
Robustness (2)			
	Format should be robust against single point of failure		
		2	Not vulnerable
		1	Vulnerable
		0	Highly vulnerable
	File corruption detection		
		2	Available
		0	Not available
	File corruption correction		
		2	Available
		0	Not available
	File format stability		
		2	Rare release of new versions
		1	Limited release of new versions
		0	Frequent release of new versions
	Backward compatibility		
		2	Large support
		1	Medium support
		0	No support
	Forward compatibility		
		2	Large support
		1	Medium support
		0	No support
	Format can only be changed via a defined procedure		
		2	Official procedure
		1	Informal procedure
		0	No procedure
Dependencies (7)			
	Not dependent on specific hardware		
		2	No dependency

Criterion (weighing factor)	Characteristic	values	
		1	Low dependency
		0	High dependency
	Not dependent on specific operating systems		
		2	No dependency
		1	Low dependency
		0	High dependency
	Not dependent on one specific reader		
		2	No dependency
		1	Low dependency
		0	High dependency
	Not dependent on other external resources		
		2	No dependency
		1	Low dependency
		0	High dependency

Appendix II

Table 2: Example application of the file format assessment method to PDF 1.4 and PDF/A

Criteria	Characteristics		PDF 1.4		PDF/A-1	
		Weight	Score	Total	Score	Total
Openness		3				
	Standardisation	7	1	2,333333[1]	2	4,666667
	Patents	7	1	2,333333	2	4,666667
	Reader with freely available source	7	2	4,666667	2	4,666667
Adoption		1				
	Market share	6	2	12	2	12
Complexity		3				
	Human readability	3	1	1	1	1
	Compression	3	1	1	1	1
	Versatility	3	0	0	1	1
Technical Protection Mechanism (DRM)		5				
	Password protection	3	1	0,6	2	1,2
	Copy protection	3	1	0,6	2	1,2
	Digital signature	3	1	0,6	2	1,2
	Printing protection	3	1	0,6	2	1,2
	Content extraction protection	3	1	0,6	2	1,2
Self-documentation		2				
	Metadata	1	2	1	2	1
	Self-describing format	1	0	0	0	0
Robustness		7				
	Format should be robust against single point of failure	2	0	0	0	0
	File corruption detection	2	0	0	0	0
	File corruption correction	2	0	0	0	0
	File format stability	2	0	0	0	0
	Backward compatibility	2	2	0,571429	2	0,571429
	Forward compatibility	2	1	0,285714	1	0,285714
	Format can only be changed via a defined procedure	2	1	0,285714	2	0,571429
Dependencies		4				
	Not dependent on specific hardware	7	2	3,5	2	3,5
	Not dependent on specific operating systems	7	2	3,5	2	3,5
	Not dependent on one specific reader	7	2	3,5	2	3,5
	Not dependent on other external resources	7	1	1,75	2	3,5
Total score				40,72619		51,42857
	Normalised to percentage of 100[2]			**70,2 %**		**88,7 %**

[1] This score is calculated as follows: PDF 1.4 receives a score of 1 because "De facto standard, specifications made available by manufacturer only" (see Appendix I). 1 is multiplied by 7 (the weight of the *Openness* criterion) divided by 3 (the number of characteristics of *Openness)* (1*7)/3=2.3

[2] The maximum score a format can receive is 58. By multiplying the total score by 100 and dividing it by 58 it is normalised to a scale from 0-100.