File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats

Carl Rauch[1]; Harald Krottmaier[2]; Klaus Tochtermann[3]

[1] Styria Media AG, Schönaugasse 64, 8010 Graz, Austria
e-mail: carl.rauch@styria.com
[2] Institute for Computer Graphics and Knowledge Visualization, Graz University of Technology
Infeldgasse 16c, 8010 Graz, Austria
e-mail: h.krottmaier@cgv.tugraz.at
[3] Knowledge Management Institute, Graz University of Technology, and Know-Center Graz
Infeldgasse 21/II, 8010 Graz, Austria
e-mail: klaus.tochtermann@tugraz.at

Abstract

While some file-formats become unreadable after short periods, others remain interpretable over a long-term. Among the over 1.000 file-formats, some are better and some are less suited for long-term preservation. A standardized process for evaluating the stability of a file-format is described in this paper and its practical use is shown with file-formats for 3D-objects. Recommendations to users of 3D-applications are given in the last section of this article. Some of the results are used in PROBADO, a sophisticated search engine for non-traditional objects (such as 3D-documents, music etc.).

Keywords: digital preservation; evaluation metric; file-formats

1 Introduction

In file-format registries like PRONOM, filext or MyFileFormat, over 1.000 file formats are registered. Even when removing all depreciated formats and even when setting the focus on one type of digital records only, e.g. 3D-objects, the number of available file formats is big (in this case among others dxf/dwg, iges, 3ds/max, 3dm, obj). While some file-formats depreciate over time, other file-formats are evolving. Formats, which were frequently used 10 years ago, are unreadable now as will many today's formats in ten years. But even slight modifications in the representation of digital objects can have major influences on their significance. An example would be a computer game with a slightly higher processing speed - it would become many times more difficult to play.

When a digital object needs to be available over a long-time period, users face the question, which file-format to choose for long-term preservation. Based on the concept of Utility Analysis [12] and on work done by Rauber, Strodl and Rauch [11], an evaluation process is described in this paper for analyzing and ranking file formats in terms of long-term reliability.

An evaluation of file-formats for 3D-objects is used for showing the process in practice. The remainder of this paper is organized as follows: Section 2 provides an overview over related work. In Section 3 the workflow and parameters for evaluating file-formats is described. In Section 4 the criteria for evaluating file-formats are shown in detail. A practical implementation for 3D-objects shows the feasibility of the described approach in Section 5.

2 Related Work

The work described in this paper is based on three research areas. The first basis is the area of digital preservation, where methods and workflows for comparing various preservation alternatives are developed and implemented. The second area are already existing initiatives to examine a file format's preservation risk. The third are file-format registries.

In the research area of digital preservation, several processes for evaluating preservation strategies were presented in the last couple of years. Among them are the test-bed workflow of the Dutch Preservation Test-bed [9] and the Utility Analysis workflow of the Vienna University of Technology [8]. As part of the DELOS

Network on Excellence project, these two workflows were combined to the DELOS digital preservation Test-bed's workflow [11], which is shown in Figure 1.

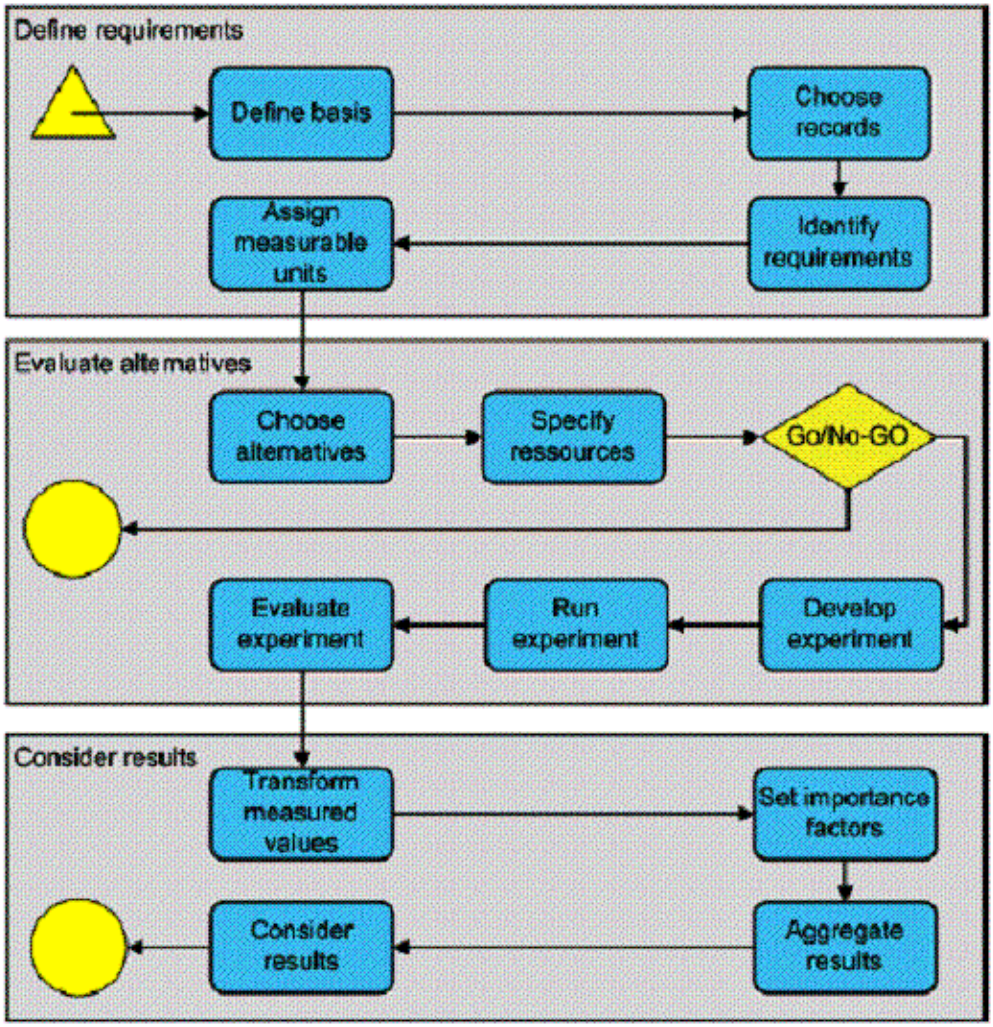

Figure 1: Overview of DELOS Digital Preservation Test-bed's workflow [11]

The DELOS workflow consists of three main parts: At the beginning, the requirements of an institution for a digital preservation strategy are defined. Here the record set, which is to be preserved, is selected, a list of criteria for evaluating the strategies defined and measurable units are assigned to each criterion. In the second part the evaluation takes place. After defining alternatives and resources to be tested, an experiment is developed and different preservation strategies are applied to the chosen objects. In the third part finally the results are examined by aggregating the performance of each alternative for the different criteria. This workflow forms the basis for the evaluation of the file-formats.

Another research is methods for evaluating the preservation risk. During the last couple of years two initiatives were started to evaluate the preservation risk of a file-format. First the INFORM system of the Online Computer Library Center [10]: There the durability of file-formats in a specific environment is evaluated, considering not only the reliability of a file-format itself, but also of the opening software, the hardware, of associated organizations, the digital archive and migration and derivative-based preservation plans. The main disadvantages of this system are, that for the assignment of a risk-factor to one of the six risk-areas, a high level of expertise is required for each individual environment. Thus the process needs highly qualified officers and cannot be standardized easily. The here proposed workflow suggests an alternative solution to these drawbacks.

A second initiative is the 'Virtual Remote Control' project of the Cornell University [4]. VRC focuses on the preservation of web pages. If the VRC-web-crawler detects a page with dysfunctional hyperlinks, longer downtimes or older server-software, the VRC-administrator is notified about the preservation risk of the web

page. VRC provides some interesting insights on evaluating the preservation risk, however it is only focusing on web pages and the file-format itself plays a minor role.

The last research area on which this paper is based is file-format repositories. Several repositories exist, where different aspects of file-formats are stored. The best-known example is the PRONOM-database of the UK National Archives. In this archive the following information are stored (among others) about a file-format [7]:

- Name, Version and other Names
- Identifiers
- Family, Classification and Orientation
- Byte Order and Related File-Formats
- Release date and support end date

A second file-format registry is FILExt. In FILExt [3] the external and internal signatures of a file-format, the software programs able to interpret the format, the MIME types, the main producing company, the file-formats name and a description is given for each file-format.

Neither of the registries contains a specific measure on the reliability of a file-format. For both the information given needs to be interpreted by a file-format expert to evaluate the appropriateness of a format for digital preservation.

3 The File-Format Evaluation Process

Based on the workflow shown in Figure 1 a process for evaluating the reliability of file-formats is presented in this section. Due to the smaller scope - the DELOS workflow is designed for comparing whole preservation strategies including appearance, process characteristics and costs - the here shown process consists of less steps than the DELOS workflow. Most of these steps are standardized for all file-formats.

1. Review Requirements: The requirements for a reliable file format are structured in a criteria-tree. The criterion focuses on two areas: on technical characteristics and on the integration of the format within the marketplace. The criteria tree described in detail in Section 4 is the same for all file-formats in order to allow comparability;
2. Assign measurable categories: The second step is to assign measurable categories to each criterion. A metric is defined describing, how to convert the measured numbers into a zero-to-five scale (e.g. number of users between 10.000 and 100.000 is equal to '3' for the market penetration criterion). These conversion tables are described in more detail in Section 4 and are standardized for every evaluation run;
3. Choose alternatives: In this step file-formats are chosen, which are evaluated during a session of the workflow. In the here presented work, six file-formats for 3D-objects are evaluated as a proof-of-concept;
4. Evaluate file formats and transform values: Based on the seven sub-criteria of the criteria tree and on the measurable categories the file-formats are evaluated and a value between zero and five (five is the best) is assigned to every criterion of each file-format. These evaluation results do typically not change over time and are stored as a basis for the final aggregation;
5. Set importance factors: After the evaluation, each criterion is ranked with a percentage value according to the user's priorities; the sum of all percentages has to be 100 %. Each user can determine the importance of certain criteria for individual circumstances with values from 0 % (is not interesting at all) to 100 % (is the only relevant criterion);
6. Aggregate results: A final value per file-format is found by multiplying the value per criterion with its weight and summing these values up. The higher the value, the better a file-format is suited for long-term preservation. By aggregating the final values of several file-formats or by taking earlier evaluations as a reference, a clear ranking can be created. A measure suggested for file-formats is the preservation risk, which is calculated by dividing the final value per file-format by the maximum value possible (in the here described metric, the maximum possible number is five). This fulfillment percentage-value has then to be subtracted from one. The higher the preservation risk, the lower the probability of being able to interpret the file-format after a couple of years.

From the above listed steps, the requirement review and the assignment of measurable categories is standardized for every evaluation run. When evaluating file-formats, a user has to do the steps three to five for each run; the aggregation of results follows again a standardized scheme.

4 The File-Format Evaluation Tree

In this section the tree of requirements and the assignment of measurable categories are described. In order to compare and evaluate file-formats in terms of long-term reliability, criteria were defined and structured in a criteria-tree. The tree is based on a discussion process with the Department for Software Technology, Vienna University of Technology, the Austrian National Library, the Austrian Phonogrammarchiv and the Dutch Nationaal Archief; it is structuring all criteria, which are seen as important to measure the long-term reliability of a file-format. The tree is shown in Figure 2. The tree consists of two branches, the technical and the market characteristics.

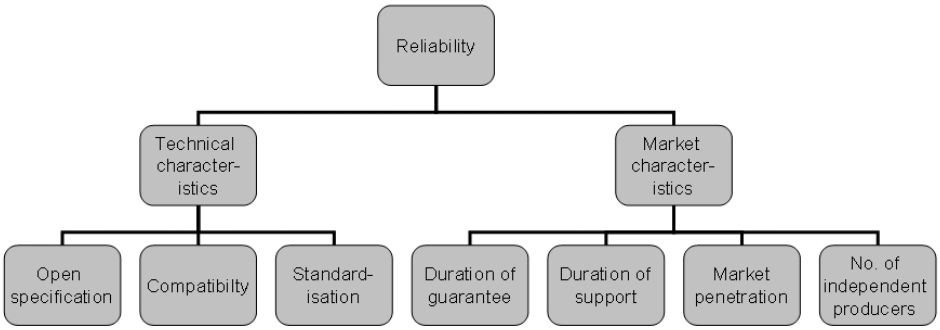

Figure 2: Criteria-tree for evaluating the long-term reliability of file-formats

The technical characteristics focus on the specification of a file-format. It consists of the following three sub-criteria:

- Open Specification: Is the specification of the file-format publicly available?
- Compatibility: Is the file-format supported and maintained by one or several software companies?
- Standardization: Is the file-format standardized by a recognized standardization agency, such as DIN or ISO?

The market characteristics focus on the acceptance and position of the file-format in the market. It is divided into the following sub-questions:

- Duration of guarantee: How long does the main producing software company guarantee to repair bugs in the interpreting software?
- Duration of support: How long does the main producing software company supports the interpreting of the file-format with its software?
- Market penetration: How many users are working with the file-format at the current time?
- Number of independent producers: How many software products exist, which are able to interpret the file-format?

In order to transform measurable units into values from zero to five the following transformation tables are suggested. The intervals are chosen in a way, which should bring a maximum distinction between typical software formats. By targeting the range of values, which typical software formats have, the differences between formats can be shown explicitly:

- Open Specification: Yes = 5, Partly available = 3, No = 0
- Compatibility: Number of software systems compatible with the format: 1 system = 1, 2 systems = 2, 3 systems = 3, 4 systems = 4, > 4 systems = 5
- Standardization: Yes = 5, Partly standardized = 3, No = 0
- Duration of guarantee: 0 years = 0 (no guarantee), > 0 years and <= 1 year = 1, > 1 year and <= 3 years = 2, > 3 years and <= 5 years = 3, > 5 years and <= 10 years = 4, > 10 years = 5
- Duration of support: 0 years = 0 (no support), > 0 years and <= 1 year = 1, > 1 year and <= 3 years = 2, > 3 years and <= 5 years = 3, > 5 years and <= 10 years = 4, > 10 years = 5
- Market penetration: < 100 users = 0, > 100 users and <= 10.000 users = 1, > 10.000 users and <= 100.000 users = 2, > 100.000 users and <= 1.000.000 users = 3, > 1.000.000 users and <= 10.000.000 users = 4, > 10.000.000 users = 5
- Number of independent producers (that support the software): 0 producer = 0, 1 producer = 1, 2 producers = 2, 3 producers = 3, 4 producers = 4, > 4 producers = 5

5 Evaluating Digital Objects for 3D-Data

As a proof-of-concept, file-formats for 3D-objects were evaluated and ranked according to their preservation risk. The steps three to six of the evaluation process are described in detail in this section.

The choice of alternatives is the first step, which needs to be done before an evaluation run. The following file-formats were selected, based on inputs from the PROBADO project [6]: Drawing Exchange Format DXF/DWG, Initial Graphics Exchange Specification IGES, 3D Studio 3DS/MAX, 3D Model 3DM and Object OBJ .

Based on publicly available sources, such as Internet queries and producer information, the file-formats were evaluated. Please note that the proof-of-concept is primarily done to show the functionality of the evaluation process and can not be seen as a final judgement on the performance of every file-format.

Criterion	DXF/DWG	IGES	3DS/MAX	3DM	OBJ
Open Specification	5	5	3	0	5
Compatibility	5	5	5	5	5
Standardization	0	5	0	0	0
Duration of guarantee	0	0	0	0	0
Duration of support	0	0	0	0	0
Market penetration	3	1	5	1	3
No. of independent producers	1	5	1	1	5

Table 1: Evaluation results per file-format

Some of the results are exemplarily described in more detail to clarify the evaluation process:

- Duration of guarantee / duration of support: No information was publicly available for these two criteria, so these criteria are always evaluated with zero (since all file-formats have the same value here, the ranking is not influenced). Data like these are typically given by software companies during sales negotiations.
- Open specification: Open specifications exist for the DXF/DWG [1], IGES [5] and 3DS/MAX [2] file-format. 3DS only gets three points, since the last found specification is from 1997, although 3DS is still under development by Autodesk.
- Compatibility of IGES: At the time of its creation IGES was compatible with most available software products. Meanwhile in PRONOM only one compatible software is listed: Adobe FrameMaker 2002; in a web-search additional software products, such as ModelPress Desktop, CrtlView or 3D Shop ModelScan are named (see http://www.programurl.com/, Date of Download: 09.03.2007). Additionally a conversion tool for Autodesk exists.
- Standardization of IGES: IGES has been standardized by the Department of Defense and the National Institute of Standards and Technology [5].
- Market penetration of 3DS MAX: Wikipedia [13] lists 42 software companies, which use the 3DS MAX format, among them major producer of computer games and animated movies.
- No. of independent producers of OBJ: According to Wikipedia, the OBJ file-format has been adopted by several software vendors and can be imported and exported to a number of software programs.

As can be seen, the above shown evaluations rely on Internet-sources only. We recommend a detailed clarification with software vendors before deciding for one format or another.

Rank	File-Format	Preservation Risk
1	IGES	40.00 %
2	OBJ	48.57 %
3	DXF	60.00 %
4	3DS	60.00 %
5	3DM	80.00 %

Table 2: The final evaluation result

After the evaluation step importance factors are set for each criterion. These factors indicate how the end-user values certain criteria. In the here shown example, all seven criteria get the same weight – 14.29 %. The evaluation results are multiplied with the weight of its criterion and summed up per file-format. By taking the percentage value from the maximum possible value (which is five) and by subtracting it from 100, the preservation risk can be obtained. The final result is shown in Table 2. The differences between the file-formats in terms of preservation risk are significant and IGES is ranked top as a format for long-term preservation.

6 Conclusion

In this paper a methodology for evaluating file-formats in terms of reliability for long-term preservation is presented. In the first part the steps of the evaluation process are described in detail. In the second part of the paper a proof-of-concept is done for 3D-file-formats to show the functionality and details of the process in practice.

After evaluating several file-formats, a file-format list can be created, where all selected formats are ranked according to their preservation risk. Such a list could be maintained by a research institution or a library and could be continually updated. By including software companies and the open-source community into the evaluation process, the evaluation results can on the one hand become more precise and on the other hand become a motivation for improving the preservation reliability of file-formats. Additionally such a ranking could be added to existing file-format registries, such as PRONOM or the Global Digital Format Repository.

Notes and References

[1] AutoCAD2006. *DXF Reference*, July 2005. URL http://www.autodesk.com/, Date of Download: 04.02.2006.

[2] Autodesk Ltd. *3D-Studio File Format*, January 1997. URL http://www.martinreddy.net/gfx/3d/3DS.spec, Date of Download: 04.02.2006.

[3] *FILExt - The File Extension Source*, 2007. URL http://FILExt.com, Date of Download: 31.01.2007.

[4] MCGOVERN, N. Y.; KENNEY, A. R.; ENTLICH, R.; KEHOE, W. R.; BUCKLEY, E. *Virtual Remote Control, Building a preservation risk management toolbox for web resources*. D-Lib Magazine Volume 10, Number 4 (2004).

[5] National Institute of Standards and Technology. Initial *Graphics Exchange Specification (IGES)*, April 1996. FIPS PUB 177-1.

[6] *PROBADO - Prototypischer Betrieb fuer Allgemeine Dokumente*, 2007. URL http://www.probado.de, Date of Download: 05.02.2007.

[7] *PRONOM, the technical registry*. URL http://www.nationalarchives.gov.uk/pronom/default.htm, Date of Download: 07.07.2006.

[8] RAUCH, C.; RAUBER, A. *Preserving digital media: Towards a preservation solution evaluation metric*. In Proceedings of the 7th International Conference on Asian Digital Libraries, Shanghai, ICADL 2004 (December 2004), Springer-Verlag Berlin, Germany, pp. 203–212.

[9] SLATS, J.; VERDEGEM, R. *Practical experiences of the Dutch Digital Preservation Testbed*. VINE, The journal of information and knowledge management systems, Volume 34, Number 2 (2004), 56–65.

[10] STANESCU, A. *Assessing the durability of formats in a digital preservation environment*. D-Lib Magazine 10, 11 (2004). URL http://www.dlib.org, Date of Download: 14.03.2005.

[11] STRODL, S.; RAUBER, A.; RAUCH, C.; HOFMAN, H.; DEBOLE, F.; AMATO, G. *The DELOS testbed for choosing a digital preservation strategy*. In Proceedings of the International Conference on Asian Digital Libraries, ICADL (2006), Springer-Verlag, Berlin, Germany.

[12] WEIRICH, P. Decision Space: *Multidimensional Utility Analysis*. Cambridge University Press, 2001. URL http://www.missouri.edu/weirichp, Date of Download: 03.08.2005.

[13] *WIKIPEDIA, The free Encyclopedia*, 2007. URL http://en.wikipedia.org, Date of Download: 20.02.2007.