Koninklijke Bibliotheek
National library of the Netherlands

Evaluating File Formats for Long-term Preservation

Caroline van Wijk / Judith Rog,

12 October 2007, iPRES 2007

Overview presentation

- Introduction National Library of the Netherlands (KB), digital archive e-Depot
- KB preservation policy, planning and strategies
- File format evaluation method
- Examples application of method
- Conclusion and discussion

National Library of the Netherlands (KB)

- Founded in 1798
- 373 employees (december 2006)
- 3,5 million ‘paper’ titles
- 10 million electronic publications
- Funded by the Ministry of Education, Culture and Science
- Deposit library since 1974; voluntary deposit, over 90% coverage

Digital archive: e-Depot

- Electronic version traditional depository
- Developed in collaboration with IBM
- Technical heart: DIAS based on OAIS
- Integrated with other library modules
- Operational since March 17, 2003
- Main archival content: journal articles

Content e-Depot

- Publications from large (inter)national publishers
- Articles mainly in PDF 1.0 to PDF 1.6

- New projects, heterogeneous content:
  - Digital master files from digitisation projects
  - Publications from university repositories
  - Websites from web archiving project

Preservation policy, planning and strategies

- KB archives publication (no editing necessary)
- Original files always kept
- Look & feel important
- Future users; render publication in ‘original environment’ and in current format
- Strategies; emulation and migration
- Choice file format; sustainability of the digital file

File Format Evaluation Method

- Quantifiable: compare formats
- Evaluation method used for digitisation guidelines, publication guidelines
- Based on sustainability criteria from DP literature
- Each criterion is broken down in characteristics
- Criteria and characteristics are weighed
- Characteristics are assigned a value; threat to long-term preservation or not?

File Format Evaluation Method - Criteria

- Openness
- Adoption
- Complexity
- Technical protection mechanism (DRM)
- Self-documentation
- Robustness
- Dependencies

File Format Evaluation Method - Characteristics

Openness:

- Standardisation
- Patents
- Reader with freely available source

Dependencies:

- Not dependent on specific hardware
- Not dependent on specific OS
- Not dependent on one specific reader
- Not dependent on other external resources

File Format Evaluation Method - Quantification

- Weighing: not all characteristics same importance
- Weighing: range from 1 to 7; 7 is most important
- Values assigned: range from 0 to 2; 2 stands for best suitable for long-term preservation
- Final score ranges from 0 to 100; 100 stands for best suitable for long-term preservation

File Format Evaluation Method – Importance

Openness	**24%**
Dependencies	**24%**
Adoption	**21%**
Complexity	10%
DRM	10%
Robustness	7%
Self documentation	4%

File Format Evaluation Method – PDF/A-1

Criterion 'Openness'

Characteristics	Weight	Value scores	Score
Standardisation	7	2	(7*2)/3=4.67
Patents	7	2	(7*2)/3=4.67
Reader with freely available source	7	2	(7*2)/3=4.67
Total score			14

File Format Evaluation Method – PDF/A-1

Criteria	Weight	Value scores	Total score
Openness (3)	24%	3*2	14
Adoption (1)	21%	1*2	12
Complexity (3)	10%	3*1	3
DRM (5)	10%	5*1,2	6
Self-documentation (2)	3,5%	1*2 + 1*0	1
Robustness (7)	6,9%	4*0 + 2*2 + 1*1	1,5
Dependencies (4)	24%	4*2	14
Normalised Score			**89**

File Format Evaluation Method – MS Word 2003

Criterion 'Openness'

Characteristics	Weight	Value scores	Total Score
Standardisation	7	0,5	(7*0,5)/3=1.67
Patents	7	0	(7*0)/3=0
Reader with freely available source	7	0	(7*0)/3=0
Total			1,67

File Format Evaluation Method – MS Word 2003

Criteria	Weight	Score	Total
Openness (3)	24%	2*0 + 1*0,5	1,7
Adoption (1)	21%	1*2	12
Complexity (3)	10%	2*0 + 1*2	2
DRM (5)	10%	2*2 + 3*1	4,2
Self-documentation (2)	3,5%	1*0 + 1*2	1
Robustness (7)	6,9%	4*2 + 3*1	3,1
Dependencies (4)	24%	3*0 + 1*1	1,8
Normalised Score			**44**

Conclusion and Discussion

- File format assessment for long-term preservation; quantifiable suitability (0-100)
- File formats can easily be compared
- Importance of criteria and characteristics transparent
- File format evaluation method essential for well thought-out choices

Conclusion and Discussion

- Are the used criteria comprehensive?
- Do the broken down criteria provide workable file format characteristic options?
- Weighing and local policy; discussion and comparison

Thank you for your attention!

caroline.vanwijk@kb.nl
www.kb.nl