A decision support system to facilitate file format selection for digital preservation

Roman Graf, roman.graf@ait.ac.at
AIT Austrian Institute of Technology, Austria
Sergiu Gordea, sergiu.gordea@ait.ac.at
AIT Austrian Institute of Technology, Austria
Heather M. Ryan, heather.m.ryan@du.edu
University of Denver, Library and Information Science Program, US
Tibaut Houzanme, houzanme@gmail.com
Indiana Archives and Records Administration, US

Libellarium, IX, 2 (2016): 267 – 274.
UDK: 025.8:007: 004.82=111
DOI: http://dx.doi.org/10.15291/libellarium.v9i2.285
Conceptual paper

Abstract

This paper presents a method to facilitate decision-making for the preservation of digital content in libraries and archives using institutional risk profiles that highlight endangered files formats (in danger of becoming inaccessible or unusable). The primary contribution of this work is the combined use of both machine-mined data and human-expert input to select and configure institution-specific preservation risk profiles. The machine-mined data used the developed File Format Metadata Aggregator (FFMA), and the crowdsourced expert input was collected via two surveys of digital preservation practitioners. A by-product of this endeavor is the ability to visualize risk factors for analysis. The underlying decision support system used the Cosine Similarity algorithm to provide recommendations for matching risk profiles to selected institutional risk settings. This method improves the interpretability of risk factor values and the quality of a digital preservation process. The aggregated information about the risk factors is presented as a multidimensional vector that shows a particular analysis focus and its resulting impact on selected file formats. Sample risk profile calculations and the visualization of risk factor dimensions are shared in the evaluation section.

KEYWORDS: digital preservation, file format, institutional risk profiles, decision support system, information aggregation

Introduction

In recent decades, libraries, archives, and museums have created digital collections that comprise millions of objects, with the goal of providing long-term access to them. One of the core preservation activities deals with the evaluation of appropriate formats used for encoding digital content. The preservation risks for a particular file format are often difficult to estimate according to existing literature (Graf, Gordea and Ryan 2014). The definition of risk factors and associated metrics is still an open research topic in the digital preservation community. Involvement of digital preservation experts is required for collecting complete information and evaluating preservation risks (Ayris et al. 2008). An example of such risk could be digital data stored in outdated or proprietary format that could not be rendered.

Societal risk factors and heuristic approach. That format obsolescence causes old digital artifacts to no longer be functional with modern software or hardware is a fact of life. Jeff Rothenberg famously mentioned in the 90's that *"Old bit streams never die—they just become unreadable [...] digital information lasts forever—or five years, whichever comes first"* (Rotenberg 1999). In that, obsolescence is both a matter of changing technology (hardware and software) and the speed with which it changes, that render older digital objects inaccessible and unusable. A proposed solution needs to address these two aspects of time efficiency and resistance to technological changes.

Current practice and perspectives. Currently, individual institutions select their own file formats for long-term preservation depending on particular projects, preservation goals, workflows, and assets. These choices reflect each organization's comfort level with risk taking/mitigation and its ability to provide for resources to support the "lifestyle" it decides for its digital objects. And even with our suggested method, it appears necessary for the institution to intervene and make decisions more rapidly about the objects' preferred formats and the importance or weight it would like to assign to each. Due to the scale of digital information that has to be managed, memory institutions need automated, or semi-automated solutions for data management and digital preservation, and the idea of an application profile for each institution would help reduce that burden by providing speed to the analysis, risk setting and free up time for format migration processing.

Proposed solution. To address these problems faster than currently afforded to the preservation community, the authors employ the FFMA system and an information integration approach (Graf and Gordea 2012). Another important open issue addressed is the risk factor estimation and customization to comply with the institutional risk profiles. The institutional risk profile can be derived from policies defined by decision-maker within the institutional context. The setup of a risk profile implies definition of institutional risk factors such as compression, storage space, lifetime, complexity, availability online, metadata

and community support, specification quality, technical protection mechanism, legal restrictions, costs, standardization, technical dependencies. The novelty of this technical solution is the employment of data mining methods to facilitate complex risk factor analysis and settings by preservation practitioners and the combining of these methods with automatically aggregated format metadata. Decision support based on the elaborated rule engine provided by FFMA and an expert knowledge base is designed to support institutions like libraries and archives with assessment for analyzing their digital assets and appropriate file formats used for encoding it. The factors utilized here for the risk metrics were provided through two studies conducted by Heather Ryan (Ryan, Graf and Sergiu 2015).

An approach to the solution implementation. The proposed mechanism would work in a similar manner as preparing institutional data curation profiles (Witt et al. 2009) or profiles set in the Library of Congress Bagger[1] for digital accessions, through the focus on evaluating and mitigating format risks. For practical considerations, generic categories of documents such as plain text, structured text, structured data and spreadsheets, presentations, databases, audio, video, raster images, vector images, raw images, geospatial data, e-mail, web pages and social media content, software, disk images, etc. (Corridan and Houzanme 2016) will each be assigned a preferred list of formats output to be migrated to. With a risk factor associated with each source and destination format, each institution can decide on their level of accepted risk related to the open preservation formats they will use to further customize their profile and stay true to their mission.

Method

File format and institutional risk profile definition system

The file format and institutional risk profile definition system (Fig. 1) shows the general workflow for the endangerment analysis of an institutional risk profile and format selection. The data for risk factor calculation are aggregated from the expert knowledge base (EKB).

Expert Knowledge Base use. The EKB employs the computed risk factors and the institutional requirements for the institutional risk profile computation. The EKB also includes information about the most often used file formats as well

1 The Library of Congress. Digital Accession Metadata Profile for use in Bagger (Json files): https://github.com/LibraryOfCongress/bagger/tree/master/bagger-business/src/main/resources/gov/loc/repository/bagger/profiles. More details can be found in *Bagger's Enhancement for Digital Accessions* (LOC blog) and *Indiana Archives and Records Administration's Accession Profile Use in Bagger* (SAA blog).

as information about "direct cause" risk factors defined by practitioners, which determine that format is endangered.

Risk representation and mitigation. Each risk profile is represented by a multidimensional vector. In the current model 31 dimensions are taken into account. A preservation practitioner may additionally estimate each of the computed risk factors in their institutional profile. When the risk level of the evaluated risk profile is too high, the collection manager can reset the institution requirements and evaluate new profile by adjusting particular risk factor values to reduce overall risk level. The computed risk profile is applied with metadata information from the FFMA, which is automatically retrieved from the Linked Open Data repositories, such as Wikidata, Pronom, and Fileinfo. While these data sources are incomplete, they provide a sufficient amount of information for initial analysis. Additional data will be collected in accompanying development of a file format endangerment index. Using the generated risk assessment data, the collection manager can make an informed decision about preservation formats.

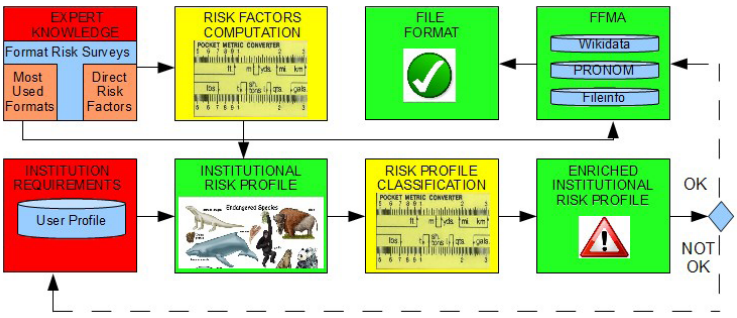

Figure 1. The overall workflow for the selection of the institutional file format

Evaluation

This evaluation aims at leveraging the domain expert knowledge base to detect the nearest risk profile as described in the workflow for auto completion of a user risk profile (see Fig. 2).

Many file formats are properly documented, have open specifications, and are more or less supported by software vendors. The above-mentioned risk factors are used as a metrics.

Processing. The processing of aggregated data for visualization of risk factor coherences and decision making for file format selection is based on computed risk profile and FFMA input. In the „Document Format Evaluation" scenario, where the authors evaluated risk profile for document formats, they performed the sample risk profile calculation and file format selection. The assumption is that an institutional user will need to define the most important risk factors and apply them as an input to the profile recommendation tool.

Tool output. The output of the tool is a complete profile in which the missing factors are accomplished with recommendations from the nearest expert risk profile computed by the tool. This evaluation is aiming at determining the most important risk factors and their visualization. This facilitates the computation of format endangerment analysis without the need to manually specify a complete risk profile; further, it helps visualize the level of agreement between important risk factors. Thus, a preservation practitioner that uses this decision-aid can adjust required risk factor settings in order to reduce or mitigate essential preservation risks.

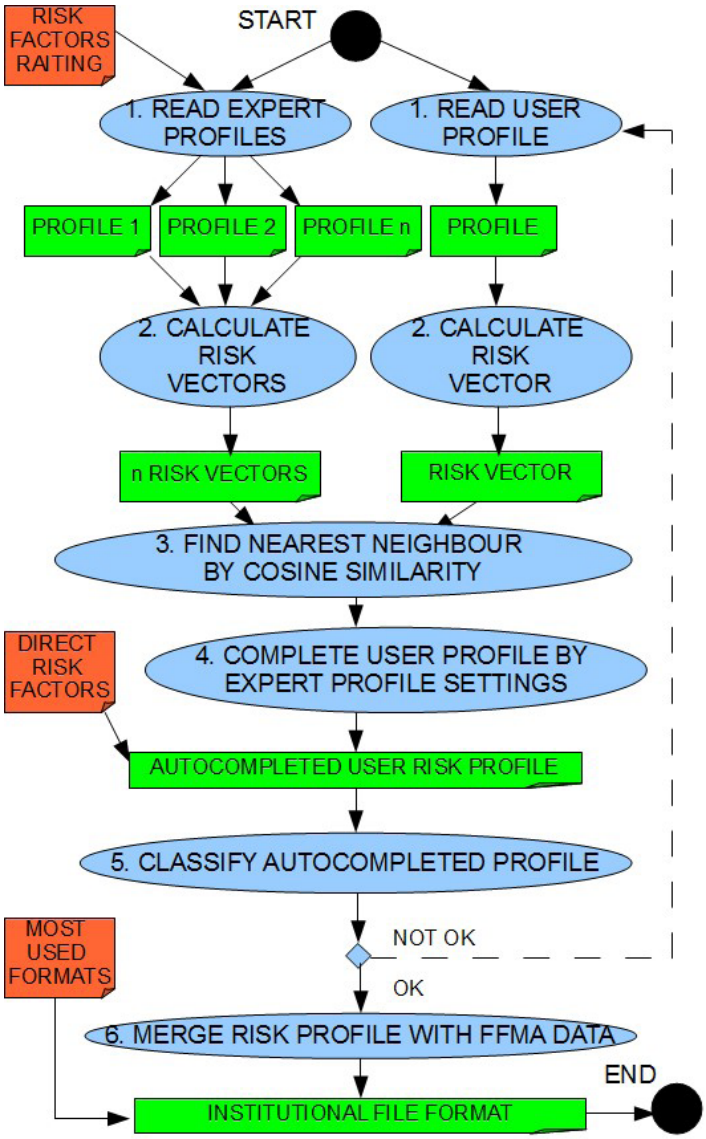

Figure 2. The workflow for recommendation of an institutional file format

Experimental results and interpretation

Figure 3 shows the visualization of the evaluated values for the most similar risk profile of „Document Format Evaluation Scenario". It also helps visualize the correlation between expert profile and one recommended profile that are vectors for 31 evaluated risk factors on the X axis. The number 31 was evaluated by domain experts during the mentioned surveys as the most important risk metrics. The Y axis is a range of the risk factor ratings. The graphic representation indicates that the selected participant profiles demonstrate good agreement with institutional profile. The institutional settings for risk factors are flagged by the green circles.

To facilitate easier evaluation, the aggregated information about the risk factors is presented as a multidimensional vector, where vector elements are values of related risk factors. E.g. for profile of expert 16 in Figure 3, geographical spread factor has rating 2, domain specificity factor has rating 1 and so far for all 31 risk factors. Rating is calculated in range from 1 to 5 and demonstrates importance of particular risk factor in a risk profile. The proposed methods will help improve the visibility of risk factor information and the quality of a digital preservation process. The well-known modified standard score method was employed to calculate similarity between risk profile vectors. This method enables selection of the most similar risk profile having defined only a few of the most important risk factors. Therefore, this facilitates complex manual comparison of multiple risk factors.

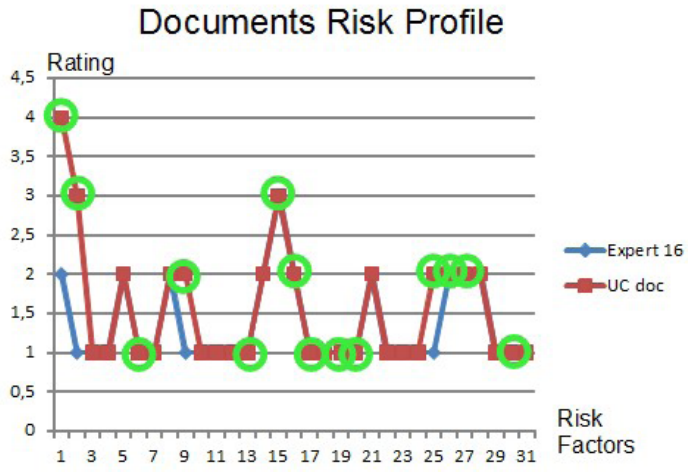

Figure 3. Plot for relation of risk factor settings for documents use case between institutional expert and the nearest expert profile

Conclusion

In this work we present an approach for the decision support for selection of appropriate file formats with the goal of assessing digital preservation processes. The presented method supports a relatively easier and faster creation of an institutional risk profile for format endangerment analysis and evaluation of file formats.

Tools and devices used. The authors made use of data mining techniques, such as the modified standard score method to analyze aggregated data and the Cosine Similarity calculation to compare risk profiles. Using the developed approach and adjusting input data, experts have the ability to choose the appropriate file format and risk factor settings for digital preservation planning in their institution. These tools and devices have been put to creative use to help solve the problem at hand.

Major contributions. The main contribution of this work is the employment of profile recommendation techniques to support risk factors set up with just a few of the most important values for a particular organization. A procedural novelty is the combination of format metadata knowledge aggregated by FFMA with institutional risk profile. The resulting risk profile is used to support digital preservation experts with semi-automatic estimation of endangerment level for file formats. A by-product of this experiment is the analysis support through the visualization and analysis of risk factors for a given set of perspectives.

Perspectives. The presented approach is designed to facilitate decision-making about preservation of digital content in libraries and archives using domain expert knowledge collected from the surveys. Further studies could focus on the proactive implementation of more automated processes that will make a greater use of computational intelligence algorithm to further make the preservation task more approachable and easier.

References

Ayris, P., R. Davies, R. McLeod, R. Miao, H. Shenton, and P. Wheatley. 2008. "The life2 final project report. Final project report." London: LIFE Project.

Corridan, J., and T. Houzanme 2016. "Selecting an integrated records and preservation management system for the Indiana Archives and Records Administration." In Building Trustworthy Digital Repositories, edited by P. Bantin. Lanham, MD: Rowman %26 Littlefield (forthcoming).

Graf, R. and S. Gordea. 2012. "Aggregating a knowledge base of file formats from linked open data." Proceedings of the 9th International Conference on Preservation of Digital Objects, poster: 292–293.

Graf, R., S. Gordea, and H. Ryan. 2014. "A model for format endangerment analysis using fuzzy logic." In Proceedings of the 11th International Conference on Digital Preservation (iPres2014), 160–168. State Library of Victoria, Melbourne, Australia.

Rotenberg, J. 1999. "Ensuring the longevity of digital information: An expanded version of the article "Ensuring the longevity of digital documents" that appeared in the January 1995 edition of Scientific American (Vol. 272, Number 1, pp. 42-7)." Accessed July 6, 2016 http://www.clir.org/pubs/archives/ensuring.pdf

Ryan, H., R. Graf, and G. Sergiu. 2015. "Human and machine-based file format endangerment notification and recommender systems development." In Proceedings of the 12th International Conference on Digital Preservation (iPres2015), Chapel Hill, North Carolina, USA, UNC.

Witt, M., J. Carlson, D. Scott Brandt, M. H. Cragin. 2009. "Constructing data curation profiles." The International Journal of Digital Curation 3, 4: 93-103. Accessed June 6, 2016. http://ijdc.net/index.php/ijdc/article/view/137/165 https://doi.org/10.2218/ijdc.v4i3.117