Human and Machine-based File Format Endangerment Notification and Recommender Systems Development

Heather Ryan
University of Denver
Denver, Colorado, USA
heather.m.ryan@du.edu

Roman Graf
Austrian Institute of Technology
Vienna, Austria
roman.graf@ait.ac.at

Sergiu Gordea
Austrian Institute of Technology
Vienna, Austria
sergiu.gordea@ait.ac.at

ABSTRACT
Effectively preserving access to digital content over time is dependent on availability of an appropriate IT infrastructure including access to appropriate rendering software and its requisite operating systems and hardware. The complexity of this task increases over time and with the size and heterogeneity of digital collections. Automating notifications on file format endangerment and decision recommendations can greatly improve preservation planning processes. This paper presents work in progress that contributes to the design and testing of an automated file format endangerment notification and recommendation system. This system's design is based on concepts explored in previous research, but it presents the novel application of statistically generated similarity profiles and machine-generated recommendations based on human expert input.

General Terms
Frameworks for digital preservation; Preservation strategies and workflows

Keywords
File format endangerment, institutional risk profiles, recommender systems, notification systems

1. INTRODUCTION
Preserving access to content encoded in particular digital file formats requires the availability of the appropriate software and hardware infrastructure. Over time, it becomes incrementally difficult to maintain this particular infrastructure and to access the stored digital content (i.e. the hardware and/or software may reach their life end). For those managing large, heterogeneous digital collections, the challenge grows with the size and variety of content aggregated in their collections. This is particularly challenging for state and government archives, which are required to preserve all content produced by their supported government agencies, regardless of format. Web archives also pose unique challenges in preservation in terms of scale and complexity.

Knowing when certain file formats are becoming endangered, meaning in danger of becoming inaccessible using commodity hardware and software; and receiving recommendations for how to maintain access to the endangered format is an important component of a sound digital preservation workflow. Having these services augmented by expert opinion and semi-automated through appropriate software support can reduce the difficulty of this challenge.

Evidence collected through an interview-based study [1] and through personal conversations with individuals managing or working with digital collections in memory institutions suggests that there is a need for systematic file format endangerment measurement, notification, and recommendation. Some indicate that such efforts are not necessary [2][3]. Arguments against these efforts cite the inherent difficulty in quantifying many file format endangerment factors and general lack of trust in automated recommender systems as inhibitors to successfully measuring file format endangerment and providing alerts and recommendations for file format risks. Other underlying concerns around developing file format endangerment measures and tools lie in expert systems' apparent circumvention of individual expertise and lack of observable data to test these measures and systems.

To reconcile these concerns, we have extended the design of the File Format Metadata Aggregator (FFMA) [4]. It now includes: 1) expert informed, hybrid decision support tools, 2) a case-based recommender system that produces recommendations according to similarity metrics [5] and initial tests on a hybrid collaborative filtering system for building/identifying institutional profiles, and 3) a knowledge based system for computing the risk factors and levels [6] on test data collected for a previous file format endangerment study [7].

The present system requires additional evaluation and testing, both through testing the system components and algorithms, and through analysis of user needs and trust in automated systems. Our first goal is to collect data on which factors digital collection managers use to assess institutional preservation friendliness. Here preservation friendliness is related to a file format's various attributes that may contribute to or hinder preservability of digital content within an institutional context. Traditionally, it has been preservation friendly formats that are selected for inclusion in digital collections managed by memory institutions [8][9][10][11][12][13]. This data will be used as corpora for testing algorithms designed to calculate institutional risk profiles. Additionally, we aim to collect information on which file formats most commonly appear in study participant collections. We use the collected list of file formats, which are sufficiently documented in Linked Open Data (LOD) repositories such as DBPedia and Freebase, as a basis for further system testing.

Our second goal is to collect information on perceived trust and utility of an automated file format endangerment notification and recommender system. Issues of trust are common complicating factors in the design and implementation of recommender systems and it is important to address them early in system design. We use information collected from this portion of the study to inform the development of additional trust-building measures such as

trustworthiness or transparency, and the ability to allow users to indicate or correct information [5].

This work in progress will lead to the novel approach to decision support for digital collection managers. While the system makes use of data-mining and statistical analysis of endangerment factors, it complements the machine learning aspects of the system with human expert input and recommendations.

This paper is structured as follows: Section 2 provides an overview of related work as well as existing work associated with this project, Section 3 explains the motivation behind each aspect of the study, Section 4 describes the study design and how it has to be applied to the design and testing of the file format endangerment notification and recommender system, and Section 4 concludes the paper and outlines planned future work.

2. RELATED WORK

This research builds on previous work on other similar efforts as well as our own related work. Similar initiatives PANIC [14], AONS II [15], SPOT [16], and the P2 Registry [17] incorporate file format identification and risk notification mechanisms.

A preliminary study has been conducted to assess file format endangerment factors [7] for measurability and fit for inclusion in a file format endangerment index. Once validated, the index will provide the framework for file format endangerment warnings. Algorithms and visualization components have been tested for risk profile definition [18] and format coherences [19].

This research improves on initial projects to extract file format data from various online resources [20] and to provide decision support using fuzzy logic [21]. Additionally, it pulls from earlier work on developing a File Format Migration Center that facilitates user-generated ratings and recommendations for file format conversion pathways [22]. The work in progress presented here builds on these previous efforts while developing novel technologies for data collection, data analysis, data visualization, alerts, and recommendation.

3. STUDY MOTIVATION

The current study is designed to contribute to the testing and further design of an automated file format endangerment notification system. Data collected is meant to be used as test corpora and to inform additional design decisions.

Initial tests of Naive Bayes analysis were performed using initial data collected for [7] which produced a successful proof of concept model for automated institutional risk profile generation. Sparse data and minimal ordinal values limited the degree to which this data set could be used for more robust testing.

Institutional risk profiles are created using human generated preference settings of institutionally-based file format evaluation factors. Recommendations for decision-making are made based on similarity calculations between the individual risk profile preferences. Similar institutional risk profiles will receive similar decision recommendations, based on expert input. It is necessary to collect more thorough input on file format evaluation factor preferences to accurately calculate the institutional risk profiles.

Previous tests of this and similar systems involved test file formats that were selected based on various criteria that may not be directly related to actual use-cases. Our intent with future system development is to test using file formats that are known to reside in digital collections currently managed in real institutional settings. To accomplish this, we are collecting information on the most commonly occurring file formats in collections managed by study participants.

Trust is a concern in the development of recommender systems, both trust in the other human contributors to the system and trust in the system's automated recommendations [7]. There are methods that can be used to ameliorate lack of trust, but presence of distrust must first be established before additional probes can be used to determine underlying reasons for extant distrust. Once reasons for user distrust are established, action can be taken to address the reasons within the design of the system.

4. STUDY DESIGN

The following study design reflects the needs outlined in the study motivation section. The study consists of an online survey administered using the Qualtrics online survey software.

4.1 Research Questions

This study is designed to answer the following research questions:

RQ1: Which factors do individuals working in libraries and archives consider to be most important when evaluating file formats for inclusion in an institution's digital collection(s)?

RQ2: Which factors do individuals consider to be causes of file format endangerment?

RQ3: To what degree do individuals working in libraries and archives believe that a file format endangerment notification and recommender system will improve their work and their preservation related decisions?

RQ4: To what degree do individuals working in libraries and archives trust the concept of an automated file format endangerment warning and recommender system?

4.2 Participants

Study participants are individuals working in libraries and archives who make decisions about digital file formats in collections they oversee. They are recruited using emails to listservs, direct email contact, and through word of mouth.

4.3 Survey Design

The survey includes four sections: Demographics, Utility and Trust, File Format Factor Rating, and Common File Formats. The study is comprised of six questions, where Question 4 contains 31 sub-questions, for a total of 36 items.

4.3.1 Section 1: Demographics

Q1. Institution type (e.g. Academic Library, City Archives, National Library, Medical Library, etc.)

4.3.2 Section 2: Utility and Trust

Q2. An application that is able to notify about endangered file formats and explain the nature of risks will improve my work and my preservations related decisions.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

Q3. I trust a computer system which is able to indicate file formats that are in danger of not being supported by commodity hardware-software systems in the near future?(10-20 years).

- Strongly agree
- Agree
- Neutral

- Disagree
- Strongly disagree

Please explain your answer.

4.3.3 Section 3: File Format Factor Rating

Q4. **(**Please rate the following 31 factors based on how important they are to consider when selecting file formats that your particular institution is able to preserve access to in the near future (10-20 years):

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

Factors:

1. **Availability Online** - the degree to which the format is available on the Web.
2. **Backward Compatibility** - whether or not newer versions of the rendering software can render files from older versions.
3. **Community Support** - the degree to which communities support the file format.
4. **Complexity** - relates to how much effort has to be put into rendering and understanding the contents of a particular file format.
5. **Compression** - whether or not, and the degree to which a file format supports compression,
6. **Cost** - The cost to maintain access to information encoded in a particular file format, e.g. to migrate files, to maintain the rendering software, or to run an emulation environment.
7. **Developer/Corporate Support** - whether or not the entity that created the original software that produces output in the file format continues to support it.
8. **Domain Specificity** - the degree to which the format is used only within specific domains.
9. **Ease of Identification** - the ease with which the file format can be identified.
10. **Ease of Validation** - the ease with which the file format can be validated, where validation is the process by which a file is checked for the degree to which it conforms to the format's specifications.
11. **Error-tolerance** - the degree to which this format is able to sustain bit corruption before it becomes unrenderable.
12. **Expertise Available** - the degree to which technological expertise is available to maintain the existence of software that can render files saved in this format.
13. **Forward Compatibility** - whether or not older versions of rendering software can render files from newer versions.
14. **Geographic Spread** - the way in which a file format is spread across the world; whether spread thinly across the globe or condensed heavily in a particular area.
15. **Institutional Policies** - the degree to which a file format is affected by institutional polices, such as whether or not an institutional policy states that content encoded in this format will be collected and preserved.
16. **Legal Restrictions** - the degree to which this file format is or can be restricted by legal strictures such as licensing, copy and intellectual property rights.
17. **Lifetime** - the length of time the file format has existed.
18. **Metadata Support** - whether or not the file format allows for the inclusion of metadata.
19. **Rendering Software Availabilty** - whether or not any type of software is available that can render the information stored in this file format.
20. **Rendering Software Functionality/Behavior Support** – the degree to which available rendering software supports various functionality and behavior encoded in a particular file format.
21. **Revision Rate** - the rate at which new versions of this file format's originating software are released.
22. **Specifications Available** - whether or not documentation is freely available that can be used to create or adapt software that can render information stored in this file format.
23. **Specification Quality** - (sub-factor of "Specifications Available") the understandability and usefulness of the format's available specifications in maintaining access to content encoded in that format.
24. **Standardization** - whether or not this file format is recognized as a standard for use and/or preservation by a reputable standards body.
25. **Storage Space** - the average amount storage space a file saved in this format requires when saved.
26. **Technical Dependencies** - the degree to which this file format depends on specific software (beyond typical rendering software), operating systems, and hardware in order for its contents to be successfully accessed or rendered.
27. **Technical Protection Mechanism** - whether or not this file format allows for or is encumbered by technical protection mechanisms such as Digital Restrictions Management (DRM).
28. **Third Party Support** - the degree to which parties beyond the original software producers support the file format.
29. **Ubiquity** - the degree to which use of this file format is widespread and in common use.
30. **Value** - the degree to which information encoded in this format is valued.
31. **Viruses** - the degree to which the format is susceptible to containing or being damaged by viruses.

The list of factors will be presented to participants in random order to enhance reliability of responses.

Q5. Which of the following factors [Backward Compatibility, Community Support, Complexity, Cost, Developer/Corporate Support, Expertise Available, Forward Compatibility, Legal Restrictions, Rendering Software Availability, Rendering Software Functionality/Behavior Support, Specifications Available, Specification Quality, Standardization, Technical Dependencies, Third Party Support, Ubiquity] do you believe is a/are direct cause(s) of file format endangerment (versus factors for evaluating whether or not a format is included in a preserved collection)?

4.3.4 Section 3: Common File Formats

Q6. Please list the most commonly appearing file formats in your institution's digital collection(s). For each file format listed:

Describe briefly their application(s) (e.g. historical photographs, institutional documents, medical records, GIS data, etc).

Explain briefly why the file format was selected for inclusion in your institution's collection(s). What advantages does it present over other, similar file formats.

5. CONCLUSION AND FUTURE WORK

The goals of this study are to inform the development of a semi-automated file format endangerment warning and recommendation system. The survey will provide insight into what participants think are the most important factors that individuals consider when evaluating file formats for inclusion in their collections. This data will serve as the test corpora for statistically determining institutional risk profiles, which will then be used to establish likeness between users. The study will also provide a use case based list of file formats that will provide a basis for realistic system experiments and tests. Lastly, the survey will help to establish the usefulness of an automated file format endangerment warning and recommender system, and to what degree people think they can trust and rely on an automated system.

Continuing research involves continued experiments and system tests, further examination of trust in automated recommender systems, and development of additional framework for system deployment and use.

6. REFERENCES

[1] Bowden, H. 2010. Assessing need for file for an automated file format obsolescence warning system for digital collections. In *iConference 2010* (Urbana-Champaign, IL, February 03 - 06, 2010).

[2] van der Knijff, J. 2013a, September 30. *Assessing file format risks: searching for Bigfoot?* Message posted to Open Planets Foundation blogs at http://www.openplanetsfoundation.org/blogs/2013-09-30-assessing-file-format-risks-searching-bigfoot

[3] van der Knijff, J. 2013b, October 8. *Measuring Bigfoot*. Message posted to Open Planets Foundation blogs http://www.openplanetsfoundation.org/blogs/2013-10-08-measuring-bigfoot

[4] Graf, R., and Gordea, S. 2012. Aggregating a knowledge base of file formats from linked open data. In *Proceedings of the 9th International Conference on Preservation of Digital Objects*, (Toronto, Canada, October 01- 05, 2012), 292–293.

[5] Ricci, F., Rokach, L., & Shapira, B. 2011. Recommender systems handbook. Springer, New York. DOI=10.1007/978-0-387-85820-3

[6] D. Heckerman. 1997. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1,1,79–119.

[7] Ryan, H. 2014. Occam's Razor and file format endangerment factors. In *Proceedings of the 11th International Conference on Digital Preservation,* (Melbourne, Australia, October 6-10, 2014).

[8] Arms, C.R., & Fleischhauer, C. 2005. Digital formats: Factors for sustainability, functionality, and quality. *Imaging Science & Technology Archiving 2005*, Washington, DC, (April 2005), 222-227

[9] Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., & Kenney, A.R. 2000. *Risk management of digital information: A File format investigation*. Washington, DC: Council on Library and Information Resources.

[10] Huc, C., et al. 2004. *Criteria for evaluating data formats in terms of their suitability for ensuring information long term preservation*. Technical Report. Groupe Pérennisation des Informations Numériques.

[11]Cornwell Management Consultants. 2005. *Selection of preservation formats: trends and issues*. Technical Report. The National Archives, U.K.

[12] InterPARES. 2007. General study 11 final report: Selecting digital file formats for long-term preservation (Version 1.1). British Columbia, Canada: McLellan, E. P.

[13] Rog, J., & Wijk, C, van. 2008. *Evaluating file formats for long-term preservation*. Technical Report. Koninklijke Bibliotheek.

[14] Hunter, J. & Choudhury, S. 2006. PANIC: An integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries, 6*(2), 174-183.

[15] Pearson, D., & Webb, C. (2008). Defining file format obsolescence: A risky journey. *International Journal of Digital Curation, 3*(1), 89-106.

[16] S. Vermaaten, B. Lavoie, and P. Caplan. 2012. Identifying threats to successful digital preservation: the spot model rsik assessment. *D-Lib Magazine*, 18, 9/10 .

[17] Carr, L., Tarrant, D., Hitchcock, S. 2011. Where the semantic web and Web 2.0 meet format risk management: P2 Registry. *International Journal of Digital Curation, 6*,1, 165-182.

[18] Graf, R., Gordea, S., Ryan, H. 2015. A Bayesian classification system for facilitating an institutional risk profile definition. In *Proceedings of the 17th International Conference on Information Technology and Engineering (ICITE*) (Oslo, Norway, July 17-18, 2015).

[19] Graf, R., Gordea, S., Ryan, H. 2015. A tool for visualization and analysis of file format coherences. In *Proceedings of the 4th International Conference of Asian Special Libraries (ICoASL)*. (Seoul, Korea, April 22-24, 2015).

[20] Graf, R., & Gordea, S. 2013. A risk analysis of file formats for preservation planning. In *Proceedings of the 10th International Conference on Preservation of Digital Objects, (Lisboa, Portugal, September 2-5, 2013)*.

[21] Graf, R., Gordea, S., Ryan, H. 2014. A model for format endangerment analysis using fuzzy logic. In *Proceedings of the 11th International Conference on Digital Preservation, (Melbourne, Australia, October 6-10, 2014)*.

[22] Bowden, H. 2009. File format migration center: Final project paper. http://longtermdata.com/docs/HBowden_ProjectPaper.pdf