# Occam’s Razor and File Format Endangerment Factors

Heather Ryan
University of Denver
Library & Information Science Program
1999 E. Evans Avenue
Denver, CO 80208
heather.m.ryan@du.edu

## ABSTRACT

Much digital preservation research has been built on the assumption that file format obsolescence poses a great risk to the continued access of digital content. In efforts to address this, a number of researchers created lists of factors that could be used to assess risks associated with digital file formats. This research examines these assumptions about file format obsolescence and file format evaluation factors with the aim of creating a simplified file format endangerment index.

This study examines file format risk under a new lens of file format endangerment. Using the Delphi method in two separate studies, this exploratory research collected expert opinion on relevance of a list of factors as causal indicators of file format endangerment.

The findings show that only three of the dozens of file format evaluation factors discussed in the literature exceeded an emergent threshold level as causes of file format endangerment: *rendering software available*, *specifications available*, and *community/3$^{rd}$ party support*. These factors are ideal candidates for use in a file format endangerment index.

## General Terms

infrastructure, communities, strategic environment, preservation strategies and workflows

## Keywords

endangerment, file formats, formative measurement model, obsolescence

## 1. INTRODUCTION

Occam’s Razor is “a scientific and philosophic rule that entities should not be multiplied unnecessarily which is interpreted as requiring that the simplest of competing theories be preferred to the more complex or that explanations of unknown phenomena be sought first in terms of known quantities” [1]. The principle of Occam’s Razor can be broadly translated into the notion that it is better to solve problems using the simplest solution.

This study, and its findings, calls into question the notion that assessing file format risk should involve complicated models with dozens of calculated and weighted evaluation factors. A conversation started by Johan van der Knijff [2][3] on the Open Planets Foundation website points out that many of the factors included in these models are theoretical, untested, and sometimes not testable. I agree.

Through the research I present here, I (and my study participants) have taken Occam’s Razor to the dozens of file format evaluation factors found in the literature. I introduce a formative measurement model, i.e., an index, as the framework to guide a more exact method of selecting a simple set of file format endangerment factors.

Within the context of this research, I also propose a shift in language usage from *obsolescence* to *endangerment*. *File format obsolescence* is a phrase commonly used to describe the phenomenon that occurs when information stored in a particular file format is no longer accessible using current technology. Although it has often been the focus of research and discussion

While the term *file format obsolescence* is still useful to describe a state in which a file format is no longer in use, I will use the term *file format endangerment* to describe the possibility that information stored in a particular file format will not be interpretable or renderable using standard methods within a certain timeframe. This term will be used in a way that is similar to its application to animal species. According to Merriam-Webster, *endanger* means, “to bring into danger or peril,” where an endangered species is “a species threatened with extinction,” or more broadly, “anyone or anything whose continued existence is threatened” [1]. A file format is not threatened with extinction or a discontinued existence; rather the threat is to the ability to access information from a file that is encoded in that format.

Using the phrase *file format endangerment* provides a new perspective for studying the nature of these risks. By studying a file format’s ability to be rendered as being similar to animal species endangerment, potentially useful parallels may be created that can lend new insight into the problem. Animal species have been studied for hundreds of years, and the methods used to document and assess the factors that contribute to their thriving or extinction can be applied to the viability or inaccessibility of the different “species” of file formats. From this we can learn which factors most heavily contribute to the risk of file format endangerment, and we can use this knowledge to identify this risk and take action to ameliorate it. Finally, the term “endangerment” embodies a sense of hope and urgency that hopefully incites action; much more so than the term obsolescence, which emits a sense of loss that is irreparable.

## 2. LITERATURE REVIEW

I explored the literature to identify and review past and present initiatives in file format risk evaluation, lists of file format evaluation factors, and measurement models that could be used to guide file format evaluation.

## 2.1 Initiatives in File Format Risk Evaluation

Several projects have approached the process of file format risk assessment and notification. These are the Automated Obsolescence Notification System (AONS), AONS II, parts of the Archive Ingest and Handling Test (AIHT), Plato, Scout, and research conducted at the Austrian Institute of Technology.

AONS[1] was a project of the National Library of Australia (NLA) and the Australian Partnership for Sustainable Repositories (APSR) and built upon work of the Preservation Architecture for New Media and Interactive Collections (PANIC) project, discussed later. In 2006, AONS was developed to create a file format obsolescence alert system, specifically for the DSpace digital repository platform. The alert system was to be built on an architecture that used DROID for file format identification, and PRONOM and Library of Congress Directory of Formats to provide obsolescence risk evaluation. If file formats found in the repository are identified to be at risk, the system generates a risk report and sends the report to the repository manager [4].

In 2007, work on AONS II began in order to refine the AONS services. Notably, the AONS II report stated, "an initial business driver for the project was a perceived need for a tool which could automate much of the assessment process, using standardized metrics that would support machine-formulation of recommendations on risk levels" [5]. Unfortunately, the project relied heavily on risk reporting capabilities of PRONOM, which have yet to come to fruition.

The Archival Ingest and Handling Test (AIHT) project[2] (2004-2005) was funded by the Library of Congress to "assess the digital preservation infrastructures of four small, real-world digital archives" [6]. The four partners were Johns Hopkins University, Sheridan Library; Harvard University Library; Old Dominion University Department of Computer Science; and Stanford University, Libraries and Academic Information Resources (Library of Congress, n.d.). As part of the AIHT, the Stanford University participants developed a file format risk-assessment system. They based their system on JHOVE for file format identification and representation information and the Arms and Fleischhauer [7] list of preferred file formats, from which they created a matrix for risk-assessment. From this they developed what they called the Empirical Walker Process, intended to be a fully automated metadata and risk-assessment generator that flags materials that may be in danger of becoming obsolete [6].

After developing this prototype system, Anderson, Frost, Hoebelheinrich, and Johnson evaluated the resources required to automate and maintain a preservation assessment of the Empirical Walker Process, such as maintaining the infrastructure to support the process. While they have yet to fully develop this process, they suggested that the cost to manage such a system was too much for one institution to bear and suggested, "perhaps a federated approach to some of this activity, as a service to a community of repositories and their users, would be most economical" [6].

Plato[3] (2005-present) was developed as part of the Planets preservation-planning project. Plato addresses many aspects of preservation planning [8]. Among them is assessing file format criteria that could indicate risk. They propose to evaluate file formats based on the criteria: browser support, standardization, ubiquity, stability, licensing, compression, format documentation, tool support, comparative file size, complexity, disclosure, master can be used as access copy, Optical Character Recognition (OCR applicable, and adoption. Becker and Rauber cite several obstacles toward realizing the goal of automating the process of measuring and evaluating formats based on these criteria: 1. only roughly 20% of the criteria can be automatically measured, 2. external sources of data or not complete and, 3. there is a lack of standardized benchmarks that can be used in comparative analysis.

Scout[4] is a semi-automatic preservation watch system being developed within the Scalable Preservation Environments (SCAPE) project (2011-present), "an EU-funded project which is directed towards long term digital preservation of large-scale and heterogeneous collections of digital-objects" [9]. Scout was designed to collect information from various sources that can be used to detect risks to digital content. It collects information from various registries like PRONOM as well as through natural language extraction from the World Wide Web [10][11]. This tool is still under development and has undergone only basic, proof-of-concept testing.

Another, similar approach toward file format risk analysis is being developed by Roman Graf and Sergiu Gordea (2011-present) [12][13], both of the Austrian Institute of Technology. They are also developing a system that collects data from various sources to analyze file formats for what they call, "preservation friendliness." They designed their system to collect data from PRONOM, DBPedia, and Freebase on twenty-one identified risk factors. They collected and analyzed data for these factors for a set of thirteen representative file formats to produce a total risk percentage value for each file format.

A few groups have developed digital preservation systems that incorporate file format risk analysis into workflows. These are the Preservation Services Architecture for New media and Interactive Collections (PANIC), Ex Libris' Rosetta, Tessella's Safety Deposit Box, and the National Library of the Netherland's (KB) *e*-Depot.

PANIC[5] (2004-2006) is a "semi-automated digital preservation system based on semantic web services" [14]. The project, funded by the Cooperative Research Centre for Enterprise Distributed Systems Technology (DSTC) and the Australian Federal Government's CRC Programme, facilitated the building of a prototype system to assess a digital object's obsolescence risk and subsequently invoke migration or emulation tools to counteract the risk. The system architecture contains invocation, notification, discover, and provider components. The invocation component was designed to detect obsolescence using information retrieved from the built-in software version registry via a notification agent. This registry contains information about software that is used to render the objects in the collection. Once notified of risk, the discovery component is set into action to locate appropriate preservation services using the OWL-S ontology that is used for describing and discovering web services. The provider component then sends the at-risk files to the located service that then performs the requested service [15]. There has been no development of PANIC beyond the prototype phase.

[1] apsr.anu.edu.au/aons

[2] www.digitalpreservation.gov/partners/aiht.html

[3] www.ifs.tuwien.ac.at/dp/plato/intro

[4] openplanets.github.io/scout

[5] www.itee.uq.edu.au/eresearch/projects/panic

Rosetta[6] (2009-present) is a digital preservation system produced by the Ex Libris Group [16][17]. The system has a deposit module, a working area, a permanent repository module, an operational repository, a preservation planning module, an administration module, and an access module. According to the software description, the preservation-planning module provides risk analysis of file formats, but there is no indication as to how this is accomplished. I contacted a representative of Ex Libris who stated that due to the proprietary nature of their product, they could not share information beyond what is available online.

Safety Deposit Box (SDB)[7] (2011-present) is part of the Preservica digital preservation suite developed by Tessella [18]. Key features of SDB are ingest, data management, storage, access, preservation planning and action, and administration. The preservation planning and action feature uses file characterization tools to assess file format risk, though there is no clear source of internal or external file format risk information and no clear evidence that this function is operational. As of this writing, the file format evaluation component of SDB is still not production ready, though, "Tessella are moving to a 'linked data' registry in the next release. The plan is to revisit the ability to define a format risk assessment in a future release once the linked data version is stable" (Evans, M., personal communication, January 24, 2014).

*e*-Depot[8] (2004-present) is a system built for the National Library of the Netherlands using the IBM system, Digital Information Archiving System (DIAS) [19]. DIAS was extended to include a Preservation Subsystem that included a functionality called the Preservation Manager that stores technical metadata that specifies the software and hardware necessary to render the file formats stored in *e*-Depot. This functionality was designed to meet three objectives: "1) Identify[ing] the electronic publications in danger of becoming inaccessible due to technology changes, 2) Planning the activities associated with preservation, i.e. implementing migration and/or emulation strategies, and 3) Specifying the software and hardware environments required to render an electronic publication" [19]. At the time of this writing, the KB web page on *e*Depot states that, "Preservation functionality will be enhanced in future DIAS versions to generate signals when stored assets must be converted or migrated to ensure their availability" [20]. Attempts to communicate with representatives from the KB to learn more yielded no results.

Digital preservation researchers and developers have put a great deal of work into creating tools and systems designed to manage and preserve digitally encoded information. A close examination of the existing tools, however, reveals a gap in a critical area of need: none of these tools and systems operationally addresses the issue of file format risk monitoring, though some developers claim their systems do or will do in the future. Many of the tools and systems discussed here claim that their file format risk analysis components will come from PRONOM, but while PRONOM has a place for it in its data model, it does not currently contain information on file format risk information. In fact, none of the tools or systems listed here has proven functionality in file format risk analysis. This shows that though the digital preservation community indicates that it is important to monitor file format risk, they have yet to find a viable way to do this.

It is not entirely clear what is preventing further progress in this area, but one obvious needed improvement is to flesh out the existing collections of file format data. Because so many of the tools and systems discussed here rely on sparse and non-existent data in the file format registries, their full functionality is hindered. Beyond this, a more clear understanding of which factors should be measured to provide proposed risk ratings will allow the community to focus its data collection efforts on the most useful and beneficial information. Before factor can be chosen and before data can be collected, it is imperative to have a clear understanding of which model to use to shape the development of a trustworthy file format endangerment measure.

## 2.2 Formative Indicators and Index Construction

Conservation biology and file format endangerment both involve the collection and analysis of data for pre-defined factors to detect potential dangers. The pre-defined factors represent indicators of the phenomenon being measured, i.e., species endangerment, epidemics, or file format endangerment; and are commonly called *formative indicators*.

Formative indicators, used in index construction, have an opposite relationship than do "effect" or "reflective indicators," which are commonly used in scale development. The opposite causal directions of reflective and formative measurement models are illustrated in Figure 1, where $\eta$ is the construct or phenomenon being measured, and $x_1$, $x_2$, and $x_3$ are the reflective and formative indicators. In panel 1, $\lambda$ represents the relationship that the construct has on the reflective indicators, $x_1$, $x_2$, and $x_3$. The symbol $\varepsilon$ represents the error. In panel 2, $\zeta$ is a disturbance term that represents remaining relationships of the construct that are not represented by the formative indicators and that cannot be measured. The symbol $\gamma$ represents the relationship that the formative indicators, $x_1$, $x_2$, and $x_3$ have on the construct and the *r* variables and their incumbent arrows represent their interdependency toward defining, creating, and indicating causes of the construct.

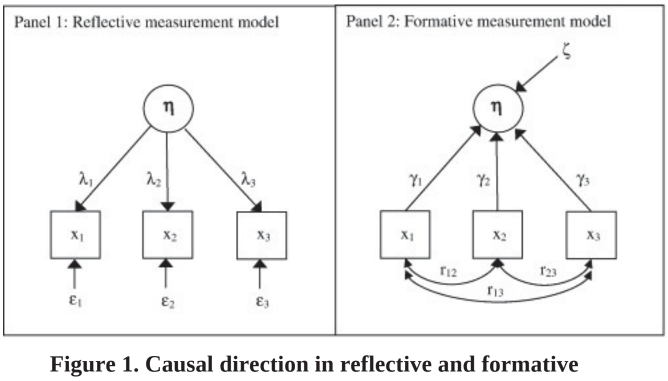
Panel 1: Reflective measurement model
η
λ1
λ2
λ3
x1
x2
x3
ε1
ε2
ε3
Panel 2: Formative measurement model
ζ
η
γ1
γ2
γ3
x1
x2
x3
r12
r23
r13

**Figure 1. Causal direction in reflective and formative measurement models [21].**

As an example of a formative measure, the construct or the phenomenon that I intend to measure is file format endangerment. The formative indicators are the factors that are determined to indicate causes of file format endangerment. In a reflective measure, the effects, i.e. the reflective indicators of the phenomenon, are measured, such as in personality measures where the personality is the construct and the personality traits are measured as an effect of the personality. According to Bollen,

[6] www.exlibrisgroup.com/category/RosettaOverview

[7] tessella.com/products/tessella-sdb

[8] www.kb.nl/en/expertise/e-depot-and-digital-preservation

"most researchers in the social sciences assume that indicators are effect indicators," where, "cause indicators are neglected despite their appropriateness in many instances" [22].

It is often not clear or obvious which of the two measurement models is most appropriate. Bollen [22] suggests that one method of determining which model is more appropriate is to perform a "temporal priority" mental experiment, or simply put, think about which happens first: the indicator or the construct. In the case of file format endangerment, my intention was to create a predictive model using factors that precede endangerment. Consequently, such a model demonstrates the temporal priority of factors that are exhibited before the phenomenon of file format endangerment. Phenomenon prediction requires data collection for *a priori* factors, or observable factors that occur before the measured phenomenon; therefore, a formative measurement model best suits the purposes of evaluating the possibility that information encoded in a particular file format will become inaccessible within a certain timeframe.

Once a researcher has determined that the indicators in question have a formative relationship with the construct, they can begin to design the measurement model, or index. Diamantopoulos and Winklehofer [23] describe the four steps for constructing an index:

1. Content Specification - defining the "domain of content the index is intended to capture"
2. Indicator Specification - choosing the indicators to be added to and tested for the index.
3. Indicator Collinearity - checking that there is not excessive collinearity between the indicators.
4. External Validity - determining that the index measures what it claims to measure and "assessing the suitability of the indicators"

Diamantopoulos and Winklehofer suggest that the definition of the domain be broad enough to encompass all of the causal indicators. Though they provide no formal recommendation for specifying which indicators to include in an index, they reported that they selected indicators for their export market sales forecasting index through "an extensive review of the forecasting literature as well as exploratory interviews with export managers" [23].

In respect to indicator collinearity, formative indictors in indexes should have a direct effect on the phenomenon being measured and have little to no intercorrelation, meaning the indicators in a formative measure should have little to no direct effect on each other. While indicators in a formative measure may have some interaction with each other, it is best if they do not have strong correlations with one another [24].

Finally, determining external validity involves testing the index to determine if it measures the specified construct. Diamantopoulos and Winklehofer suggest, "One possibility is to use as an external criterion a global item that summarizes the essence of the construct that the index purports to measure" [23].

The research presented here addresses the first two of the above steps. For the first step, I specify the content of the file format endangerment index as being all factors that indicate a cause, either through their presence or absence, information encoded in particular file formats to become inaccessible over a specified timeframe. Similar to Diamantopoulos and Winklehofer, I addressed indicator specification through an extensive literature review, supplemented by the factor-rating Delphi exercise described below. I intend to address steps three and four in future research.

## 2.3 File Format Evaluation Factors in the Literature

Effective analysis of file format endangerment requires a well-constructed and validated index to guide data collection. The key to creating a valid index is choosing the right factors that have a formative relationship with the measured phenomenon. Previously, researchers from various institutions created several different lists of file format evaluation criteria. Some of these lists of criteria were designed to evaluate aspects of file formats that can contribute to or alleviate risks associated with file formats. While none of these lists were created with the intention of creating a file format endangerment index, the approaches used are similar enough to provide a useful starting point for the index development process.

At the beginning of this research process, I identified twelve sets of file format evaluation criteria from the literature listed in Table 1. Within these lists, I identified 138 individual factors. The lists have varying numbers of factors. Some had as few as five factors, and one had as many as 22.

**Table 1. Sources of File Format Evaluation Factors**

| Project/Program/Institution | Year |
|---|---|
| Risk Management for Digital Information Project; Council on Library and Information Resources [25] | 2000 |
| Math*Diss* International Project and EMANI project; Niedersächsische Staats- und Universitätsbibliothek, Götingen [26] | 2003 |
| Groupe Pérennisation des Informations Numériques (PIN) [27] | 2004 |
| Internetbevaringsprojektet (the Internet Preservation Project); Statsbiblioteket (The State Library), Det Kongelige Bibliotek (Royal Library, Denmark) [28] [29] | 2004 |
| INvestigation of Formats based on Risk Management (INFORM) [30] | 2004 |
| Automated Preservation Assessment of Heterogeneous Digital Collections (AIHT) [31] [32] | 2005 |
| The National Archives (TNA-UK) [33] [34] | 2005 |
| Service Oriented Architecture (SOA); University of Minho, Portugal [35] [36] | 2006<br>2007 |
| International Research on Permanent Authentic Records in Electronic Systems 2 (InterPARES) [37] | 2007 |
| National Centre for Radio Astrophysics [38] | 2007 |
| Koninklijke Bibliotheek (KB) (The Royal Library, Netherlands) [39] | 2008 |
| Preservation and Long-term Access through Networked Services (PLANETS) [40] | 2008 |

## 3. RESEARCH METHOD

One of the primary objectives of this research was to clarify which of the many factors discussed in the literature are the most relevant formative indicators to include in a file format endangerment index. The research described here took a three-pronged approach to addressing these issues: two separate Delphi studies and one information gathering and rating exercise designed to test a unification of the two Delphi studies.

The Delphi method was the most effective method to determine which are the most relevant factors that indicate a cause of file format endangerment. When little data exists on a topic, such as with file format endangerment, Delphi is known to be an effective method of "producing trustworthy personal probabilities regarding hypotheses" in experts' knowledge area [41]. Dalkey [42] explained that characteristics of a Delphi procedure are anonymity, iteration with controlled feedback, and statistical group response. These procedures were designed to reduce "the influence of certain psychological factors, such as specious persuasion, the unwillingness to abandon publicly expressed opinions, and the bandwagon effect of majority opinion." Gordon and Helmer suggested that inviting participants to review other panel members' reasoning will promote a thoughtful consideration of ideas and will lead to a more accurate representation of the truth. [43]

After performing Bollen's [22] temporal priority mental experiment, described in Section 2.2, I determined that the factors I was examining for file format endangerment occurred before the phenomenon of file format endangerment. This pre-phenomenal occurrence indicates that the factors should be considered as potential causal indicators of file format endangerment, and thus appropriate for use in an index.

## 3.1 Selecting File Format Endangerment Factors for Review

My review of existing literature revealed many discussions of the importance of assessing a file format's stability for long-term preservation. Several of these discussions include proposed measures for assessing file formats for preservation purposes, as discussed in the literature review. I used these lists as the starting point for what eventually became the list of file format endangerment factors rated in the Factor Rating Questionnaire.

I used a semi-structured method to compile a draft list of factors. I copied each of the evaluation criteria into a document with citations to the original reports for reference. I then compiled all of the factors into one list, removing exact duplicates as I went. This process resulted in a list of nearly fifty factors.

I then started a new list of factors, grouping similar factors together by reviewing provided descriptions. For example, I grouped *widely accepted*, *widespread use*, *popularity*, *market share*, and *adoption* under the factor *ubiquity*. I evaluated each group of similarly themed factors and selected a name for the group that best described them. I made no value judgments as to the factors' viability as formative indicators of file format endangerment. This process resulted in a list of twenty factors. I then wrote definitions for each of the remaining factors.

I provided a list of all of the factors that were presented in the literature to a knowledgeable colleague who independently performed the same task. There were a number of differences in the way this person grouped and named the factors. We met and discussed each of our factor groupings and reached an agreement on the final synthesis of factor lists. The following are the resulting factors and their definitions:

**Backward/Forward Compatibility** - whether or not newer versions of the rendering software can render files from older versions, or whether or not older versions of rendering software can render files from newer versions.

**Community/3rd Party Support** - the degree to which communities and/or parties beyond the original software producers support the file format.

**Complexity** - relates to how much effort has to be put into rendering and understanding the contents of a particular file format.

**Compression** - whether or not, and the degree to which a file format supports compression.

**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.

**Developer/Corporate Support** - whether or not the entity that created the original software that produces output in the file format continues to support it.

**Ease of Identification** - the ease with which the file format can be identified.

**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.

**Error-tolerance** - the degree to which this format is able to sustain bit corruption before it becomes unrenderable.

**Expertise Available** - the degree to which technological expertise is available to maintain the existence of software that can render files saved in this format.

**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.

**Lifetime** - the length of time the file format has existed.

**Metadata Support** - whether or not the file format allows for the inclusion of metadata.

**Rendering Software Available** - whether or not any type of software is available that can render the information stored in this file format.

**Revision Rate** - the rate at which new versions of this file format's originating software are released.

**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.

**Standardization** - whether or not this file format is recognized as a standard for use and/or preservation by a reputable standards body.

**Storage Space** - the average amount storage space a file saved in this format requires when saved.

**Technical Dependencies** - the degree to which this file format depends on specific software, operating systems, and hardware in order for its contents to be successfully accessed or rendered.

**Technical Protection Mechanism** - whether or not this file format allows for or is encumbered by technical protection mechanisms such as Digital Restrictions Management (DRM).

**Ubiquity** - the degree to which use of this file format is widespread and in common use.

## 3.2 Research Design

This research involved the use of four questionnaires; administered online using Qualtrics survey software:

1. A questionnaire designed to collect information about the quantity and quality of experience that recruited Delphi participants had working with file formats in a digital preservation context. I used the information collected from

this questionnaire to determine the expertise level of participants and to assign them to one of the two Delphi groups.

2. A questionnaire designed to collect information on participant opinions of file format endangerment level ratings of 50 test file formats. I administered this questionnaire in a Delphi process in which participants answer the questionnaire over multiple rounds and review anonymous responses of their fellow participants between rounds.
3. A questionnaire designed to collect information on participant opinions of the relevance of factors as a cause of file format endangerment.
4. A questionnaire designed for one special rater participant to collect and report on information about factors for a list of file formats, to collect endangerment level ratings for the list of file formats, and to collect relevancy ratings for the list of factors considered as causes of file format endangerment. I designed this exercise to provide an additional source of data collection for both understanding the current perceived level of file format endangerment and for understanding which factors are direct causes of file format endangerment.

The results presented here are focused primarily on the third and fourth questionnaire. In the third questionnaire, I presented participants with the list of file format evaluation factors compiled from the dozen file format evaluation lists found in the literature.

In this questionnaire, I asked participants to rate each factor on an ordinal scale that indicates degrees of relevancy of the factor as a cause of file format endangerment:

- Not relevant at all
- Somewhat relevant
- Very relevant

I also asked participants to provide a brief narrative to explain their ratings for each of the factor options. Additionally, I asked participants to suggest factors that they believed to be a cause of file format endangerment that were not included in the original list, and their rational for suggesting the factors.

After participants completed their questionnaires, I created a document with participants' anonymized ratings and explanatory narratives for each questionnaire. I shared this document with participants and asked them to review each other's answers and narratives, and to thoughtfully reconsider their original answers. I then asked them to answer a fresh version of the questionnaire in a second round.

Some participants suggested additional index factors during the first round of the Factor Rating Questionnaire. I reviewed the 16 suggested factors, and from them, selected six new factors that had not in some way been addressed by the original list of 21 factors. For example, one participant suggested, "Existence of a community around the format," however, this factor was already addressed under the factor, *community/3$^{rd}$ party suppor*t.

Additionally, I evaluated the justification narratives in the first round of the Format Rating Questionnaire for the emergence of additional factors that should be included in the Factor Rating Questionnaire. Based on this evaluation, I added the factor, *value* to the second round of the Factor Rating Questionnaire. I added a total of seven new factors to the second round of the Factor Rating Questionnaire and asked participants to rate them on the same scale as the original twenty-one factors. The following are the seven new factors that I added to the original 21 factors to be rated in Questionnaire 2, Round 2:

**Value** - the degree to which information encoded in this format is valued.

**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.

**Domain Specificity** - the degree to which the format is used only within specific domains.

**Viruses** - the degree to which the format is susceptible to containing or being damaged by viruses.

**Availability Online** - the degree to which the format is available on the Web.

**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.

**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.

I asked participants to answer the Factor Rating Questionnaire for a third time with only the seven new factors introduced in the second round. This gave participants an opportunity to rate the new factors a second time. As with previous rounds, I collected the anonymized responses into a document and asked participants to review the document as they re-rated the factors. After the second round of rating for each factor, I determined that participant ratings had not changed substantially enough to continue to additional rounds.

The fourth questionnaire was administered to one trained, special reviewer. In this questionnaire, the reviewer was presented with each of the file formats that were used in the Format Rating Questionnaire. For each file format, I asked the reviewer to:

1. Review a guide on possible data collection sources that I created based on data I collected from the file format rating Delphi questionnaire.
2. Collect and share information from online sources, other recommended sources, or from personal knowledge for each of the factors selected during data analysis of the Factor Rating Questionnaire.

After considering the data collected in step 2, I then asked the reviewer to rate each file format on the file format endangerment level scale used in the Format Rating Questionnaire:

- Information stored in this file format is already inaccessible.
- Information stored in this file format will be inaccessible in 1-5 years.
- Information stored in this file format will be inaccessible in 6-10 years.
- Information stored in this file format will be inaccessible in 11-20 years.
- Information stored in this file format will be inaccessible in 20 years or more.
- I am not familiar enough with this file format to rate it.

After the reviewer collected factor information for each of the forty-three file formats, I asked him to rate each of the factors using the same scale for relevancy as a cause of file format endangerment that was used in the Factor Rating Questionnaire:

- Not relevant at all

- Somewhat relevant
- Very relevant

Because the special rater had just gone through the exercise of searching for information on each factor and applying this directly to rating the file formats, his ratings were strongly based in the reality of putting the factors to use in a real-world scenario. This activity provided me with additional data that I used to compare with other factor-related data that I collected from the file format rating and factor rating Delphi questionnaires.

I conducted a semi-structured e-mail interview in which I elicited feedback on the process the special reviewer used to collect information for each factor, how useful he found each factor to be in assessing file format endangerment, and any other thoughts and opinions he had about the process.

## 3.3 Participants

I selected participants for the two Delphi questionnaires from a group of individuals I identified as having expertise on file formats. Luo and Wildemuth recommended that experts be chosen based on "practical experience in implementing, managing, and evaluating [the desired expertise topic]; research experience in studying [the desired expertise topic]; publications on the topic, and so on" [44]. Based on these recommendations, I chose recruits for the Delphi questionnaires who have demonstrated experience in managing and evaluating file formats in a digital preservation environment, conducting research on file formats in digital preservation, and/or producing publications on the topic. These people have demonstrated experience in these areas either through producing publications, giving presentations, teaching workshops or courses, or writing blog posts about working with or evaluating file formats in a digital preservation context. Additionally, several people were identified as file format experts by experts already identified for the study.

Delbecq, Van de Ven, and Gustafson [45] recommended that for a homogenous group, ten to fifteen participants is adequate to form a Delphi panel. Accordingly, the aim for this study was to assemble two groups of 10-15 expert participants for the two-phase Delphi portion of the study. I initially recruited a total of 25 participants for the Delphi studies. Of these twenty-five participants, four dropped out of the study before the Delphi questionnaire process began. Twenty-one participants completed all or most of the Delphi questionnaires, with 10 participants in one study, and 11 in another.

Participants reported file format experience ranging from one to thirty years. The twenty-one participants reported a total of 210 years of working with file formats in a digital preservation context. The study includes some participants with a comparatively low number of years of relevant experience, but who are included because of the high quality of experience reported.

I recruited one additional participant to serve as a special reviewer for the fourth questionnaire of the study. This reviewer demonstrated a basic understanding of file formats and the challenges they pose to digital preservation. The reviewer demonstrated an aptitude to be trained for this study and was able to demonstrate skills in searching for information about file formats and for rating file format endangerment levels. The reviewer was trained in a one-on-one session where I reviewed the factors, the file formats, and the data collection guide that I created for him.

# 4. RESULTS

I asked expert participants to rate factors for relevancy as a cause of file format endangerment in order to make sense of the dozens of factors discussed in the literature and to elicit their views on which of the factors have a direct effect on the ability to access information encoded within a particular file format. Both the numerical ratings and participant comments provided insight into this issue.

First, the numerical ratings provided a cutoff for which factors participants believed were at least *somewhat relevant*. With the *somewhat relevant* rating having a value of 0.50, anything that received a rating below 0.50 did not make the cutoff. Half of the factors were rated at 0.50 and above. This cutoff allowed me to eliminate the half of the factors that were rated below 0.50, focusing instead on those factors that the experts deemed to be most relevant. No factor received unanimous ratings of *very relevant*.

Only six factors were rated at 1.00, which is the halfway point between *somewhat relevant* and *very relevant*. If I were selecting factors based solely on the data collected from this Delphi study, this would be the most logical cutoff point, as 1.00 is a good candidate value for a simply "relevant" rating. The factors that were rated at 1.00 and above were: *specification quality* (1.00), *expertise available* (1.05), *community/3$^{rd}$ party support* (1.05), *technical dependencies* (1.05), *rendering software available* (1.14), and s*pecifications available* (1.41).

The comments from participants provided insight into the complex nature of the issue. Many of the comments reflected the ambiguity of some of the factors. For example, one participant wrote about *complexity*, "This is an 'it depends' answer - complexity is hard to bundle into one type of characteristic. Different types of complexity could be answered on their own." Another wrote on the *cost* factor, "I agree with round 1 responses that state cost as a complex, multi-faceted and organizational[ly] influenced factor." Other factors proved to be less ambiguous and participants were able to more directly justify their ratings.

The fact that only six factors were rated at 1.00 and above is an important finding. I began this research with a total of 138 individual factors that I found in the literature. I was able to reduce this list of factors to 21 factors. Through the Delphi process, I was then able to reduce this number to six factors that participants rated as at least halfway between *somewhat relevant* and *very relevant*. Reducing the number of factors this amount was a large step toward the final selection of clear formative indicators for a file format endangerment index.

Table 2 shows a comparison of ranked factors in order of prevalence (in the case of the format rating justification text count) and rating level (Delphi factor rating means and special rater ratings). Examining each dataset included in this table reveals cutoff points for which factors are the most important for indicating file format endangerment. In the Delphi format rating justification text coding count data, there was a distinct drop-off of factor appearances after *specifications available*. While *legal restrictions* appeared in the format rating justification text 97 times, the next most frequently appearing factor, *complexity*, only appeared 63 times. This left *rendering software available*, *ubiquity*, *specifications available*, and *legal restrictions* as well-agreed-upon factors to consider in further analysis.

A logical cutoff point for both the Delphi factor rating mean ranking and special rater factor ratings datasets is a rating above 1.00, the halfway point between *somewhat relevant* and *very* *relevant*. A rating above 1.00 indicates that the factor was rated close to *very relevant*, whereas factors rated at or below 1.00 are at most *relevant*. For the Delphi factor rating mean ranking this leaves the factors *specifications available*, *rendering software available*, *technical dependencies*, and *community/3rd party support*. For the special rater factor ratings this leaves *rendering software available*, *specifications available*, *ubiquity*, and *community/3rd party support*.

**Table 2. Factor data comparison chart demonstrating cutoff points for emergent and most relevant factors**

| Delphi Format Rating Justification Text Factors (# of appearances in text) | Delphi Factor Rating Mean Ranking (mean rating value) | Special Rater Factor Ratings (mean rating value) |
|---|---|---|
| Rendering Software Available (162) | Specifications Available (1.40) | Rendering Software Available (1.50) |
| Ubiquity (130) | Rendering Software Available (1.10) | Specifications Available (1.50) |
| Specifications Available (111) | Technical Dependencies (1.10) | Ubiquity (1.50) |
| Legal Restrictions (97) | Community/3rd Party Support (1.10) | Community/3rd Party Support (1.50) |
| Complexity (63) | Expertise Available (1.00) | Legal Restrictions (0.50) |
| Community/3rd Party Support (51) | Legal Restrictions (1.00) | Technical Dependencies (0.50) |

After comparing the results from the three sets of collected data, five factors emerged as being either more highly ranked, or as appearing more times in the format-rating justification text. Examining each of the five remaining factors in light of the qualitative data collected provides more clarity for which are the most relevant as candidate causal indicators of file format endangerment.

**Rendering software available**. *Rendering software available* and *specifications available* are the only two factors that appeared beyond the cutoff point in all three datasets. It appeared as the top factor in two of the three datasets, and would have tied for the top ranking in the Delphi factor rating dataset if not for one *not relevant at all* rating. The rationale for this aberrant rating was justified that the participant considered the lack of rendering software to be the definition of obsolescence/file format endangerment and therefore rated it as being not relevant within the context of the participant's self-selected definition.

Four of the eight participants who rated this factor as *very relevant* indicated lack of rendering software strongly suggests file format obsolescence. For example, one participant wrote, "By definition without rendering software the format is obsolete." By far, the comments about the *rendering software* factor in the Delphi factor rating exercise were very strong, simple, and direct: without rendering software a file format is essentially obsolete. The strength of the comments about this factor points to it being a very strong candidate as a direct cause of file format endangerment.

**Specifications available.** Like *rendering software available*, the *specifications available* factor was included beyond the cutoff point in all three factor evaluation datasets in this study. It received a very high relevancy rating (1.40 of 1.50 possible) from the Delphi factor rating participants. Delphi participants indicated that having specifications available enables the creation of rendering software if none is available. Furthermore, others indicated that it helps to determine if software faithfully renders the contents of a file. One participant wrote, "It is hard to see that a format would not be more endangered if specifications could not be obtained." Based on the ratings and the strength of the participant comments, the *specifications available* factor is another strong candidate as a cause of file format endangerment.

**Ubiquity**. The case for considering the *ubiquity* factor as a cause of file format endangerment is weakened for several reasons. First is the fact that it only remained above the cutoff point in two of the three datasets. Second, though the special rater rated it as *very relevant*, he explained later that he only considered it to be a secondary factor, because of the following scenario: "there are also formats that are not widely distributed that are not endangered at all, such as the .nes format, used for ROM dumps of Nintendo Entertainment System cartridges."

This sentiment is echoed in many of the Delphi factor rating comments, where several participants described its effect on endangerment in secondary terms. For example, one participant wrote, "The popularity of a given file format increases the support provided by user communities and consequently increases the resources allocated/available for development/maintenance for further developments." In this scenario, the ubiquity of the file format has an effect on other factors that directly affect the endangerment level of the format and serves more as a tertiary factor that affects *community/3rd party support*.

**Community/3rd party support.** This factor is ultimately a secondary factor, even though it appeared above the cutoff point in two of the three datasets. Participants in the factor rating Delphi referred to it as a stopgap against a single point of failure: "single-point of failures are serious potential problems, and having a format which is supported by a single provider, rather enjoying larger community and 3rd party support, is a classic single point of failure situation. The wider the experience with and understanding of a format, the better, and the lack of those can present serious risks." In this case, community/3rd party support is a factor that can directly support the existence of rendering software, but is often contingent on the availability of specifications.

**Technical dependencies.** This factor appeared above the cutoff line in only the Delphi factor rating dataset. The special rater noted that he "didn't find technical dependencies to be a useful indicator as all formats have some technical dependencies." When the format rating Delphi participants mentioned technical dependencies, it was typically in the context of causing problems with the full and faithful rendering of a file that calls in information from external files; but do not mention it preventing a file from being rendered at all. In this case, *technical dependencies* is a tertiary factor where *rendering software* is the primary and *rendering software feature/functionality/behavior support* is the secondary factor.

**Legal restrictions.** This factor appeared above the cutoff line in only the Delphi format justification text coding dataset. Close examination of temporal priority reveals that while legal restrictions do have an effect on accessibility of digital content, this factor is actually a secondary factor to *specifications available* and *community/3$^{rd}$ party support*. The instances where legal restrictions were coded in the format rating justification text were those times where participants mentioned the availability of specifications and the existence of open source software. Legal restrictions can prohibit the free availability of specifications and prohibits the creation of rendering software through third parties.

It was through the process of comparing these results and scrutinizing the remaining factors that I was able to make a final reduction in factors from six to three: *rendering software available*, *specifications available*, and *community/3$^{rd}$ party support.* From beginning to end, I was able to reduce the list of factors from the original 138 factors that I found in the literature to three, for a total reduction of 135 factors.

## 5. CONTRIBUTIONS AND IMPLICATIONS

The findings of this research suggest that the three top contenders for use in a file format endangerment index are *rendering software available*, *specifications available*, and *community/3$^{rd}$ party support*. This is a marked reduction from previous total of the 21 factors synthesized from the original list of 138 factors I found in the literature. The benefit of which is that file format data collection can be focused in the areas defined in the index.

The research discussed here is the first step toward creating a file format endangerment index that can be used to detect when content encoded in a particular file format may be more difficult to access over time. Following the recommendations of Diamantopoulos and Winklehofer [23] for constructing an index, the next steps are to test and validate the index. Testing and validating an index first requires that data be collected for the selected formative indicators.

A starting point for data collection can be to use the data collected by the special rater and the data collection suggestions provided by the factor rating Delphi participants. From there the index can be validated against the file format ratings collected in the format rating Delphi study, future collected expert ratings, and other external sources. From there, continued data collection for each of the factors can be conducted in conjunction with continued assessment of the collected data.

Once the factors selected for the index have been adjusted and validated, the measure can be put to use in evaluating file format endangerment levels both in the local and global contexts. Coordination of cooperative efforts with institutions, coalitions, and other researchers who are working in this area can expand data collection and the application of the index.

Additionally, it will be valuable to explore nuances of each of the factors. For example, the factor, *specifications available*, could be examined not just by whether or not specifications are available, but by how useful the specifications are to the creation or recreation of viable rendering software. Additionally, the factor, *rendering software available*, could be evaluated not just for whether or not software is available, but how faithfully it represents the original intended representation of the encoded content.

In performing this study, I have used a hypothetical Occam's Razor to cut away what had previously been an unmanageably large collection of mostly inoperable file format endangerment factors to leave just three factors that can be used in a file format endangerment index. The simplification of factors and the creation of the file format endangerment index contributes to the digital preservation community's ability to know which file formats are at risk so issues can be addressed before they becomes too expensive and time consuming to manage in the future.

## 6. ACKNOWLEDGMENTS

I would like to thank all of the study participants. They generously donated a great deal of time to this study and provided surprisingly detailed and thoughtful responses to the questionnaires. I would also like to thank my dissertation chair and dissertation committee for their guidance and support in the conception and execution of this research.

## 7. REFERENCES

[1] Merriam-Webster. 1994. *Merriam-Webster's collegiate dictionary* (10th ed.). Springfield, MA: Merriam-Webster, Inc.

[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 blogshttp://www.openplanetsfoundation.org/blogs/2013-10-08-measuring-bigfoot

[4] Curtis, J. 2006. *AONS system documentation* (Revision 169 2006-09-29). Technical Report. Australian Partnership for Sustainable Repositories.

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

[6] Anderson, R., Frost, H., Hoebelheinrich, N., & Johnson, K. 2005. The AIHT at Stanford University. *D-Lib Magazine*, 11,12.

[7] 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.

[8] Becker, C., & Rauber, A. 2011. Decision criteria in digital preservation: What to measure and how. *Journal of the American Society for Information Science and Technology*, 62, 6, 1009-1028.

[9] Scalable Preservation Environments [SCAPE]. n.a. *About SCAPE*. Retrieved December 29, 2013 from http://www.scape-project.eu/about

[10] Faria, L. 2013. *Scout - A preservation watch system*. Retrieved December 29, 2013 from the Open Planets Foundation website http://www.openplanetsfoundation.org/blogs/2013-12-16-scout-preservation-watch-system

[11] Faria, L., Akbik, A., Sierman, B., Ras, M., Ferreira, M., & Ramalho, J.C. 2013. Automatic preservation watch using extraction on the web. *Proceedings of the10$^{th}$ International Conference on the Preservation of Digital Objects, Lisbon, Portugal*.

[12] Graf, R. & Gordea, S. 2013. A risk analysis of file formats for preservation planning. *Proceedings of the10$^{th}$ International Conference on the Preservation of Digital Objects, Lisbon, Portugal*.

[13] Graf, R., Gordea, S., & Ryan, H. 2014. A model for format endangerment analysis using fuzzy logic. *Proceedings of the11th International Conference on the Preservation of Digital Objects (iPres2014) (accepted for publication), Melbourne, Australia*.

[14] Hunter, J. & Choudhury, S. 2006. PANIC: an integrated approach to the preservation of composite digital objects using semantic web services. *International Journal of Digital Libraries*, 6,2, 174-183.

[15] Hunter, J. & Choudhury, S. 2004. A semi-automated digital preservation system based on semantic web services. In *Global Reach and Diverse Impact: Fourth ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '04), June 7-11, 2011,Tucson, AZ, pp. 268-278. Association for Computing Machinery.

[16] Ex Libris Group. 2010. *Ex Libris Rosetta: A digital preservation system product description.* Retrieved February 21, 2013 from http://www.exlibrisgroup.com/category/RosettaOverview

[17] Peled, I. 2011. The challenges of building Ex Libris Rosetta, a digital preservation system. *Liber Quarterly*, 21, 1. http://liber.library.uu.nl/index.php/lq/article/view/8012/8354

[18] Tilbury, J. 2014. *The active preservation of digital information*. White Paper. Tessella Group.

[19] Oltmans, E., van Diessen, R.J., & van Wijngaarden, H. 2004. Preservation functionality in a digital archive. In *JCDL 2004: Proceedings of the Fourth Acm/Ieee Joint Conference on Digital Libraries: Global Reach and Diverse Impact: Tucson, Arizona, June 7-11, 2004,* edited by Hsinchun Chen, Michael Christel and Ee-Peng Lim, 279-86. New York, NY: ACM Press.

[20] Koninklijke Bibliotheek. n.d.. *More about the e-Depot*. Retrieved June 7, 2013 from http://www.kb.nl/en/expertise/e-depot-and-digital-preservation/more-about-the-e-depot.

[21] Diamantopoulos, A., Riefler, P., & Roth, K.P. 2008. Advancing formative measurement models. *Journal of Business Research,* 61, 1203-1218. P. 1205

[22] Bollen, K. A. 1989. *Structural equations with latent variables*. New York: Wiley-Interscience.

[23] Diamantopoulos, A. & Winklhofer, H.M. 2001. Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38(2).

[24] Petter, S., Straub, D., & Rai, A. 2007. Specifying formative constructs in information systems research. *MIS Quarterly, 31*,4, 623-656.

[25] 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.

[26] Fischer, T. 2003. LaTeX as an archiving format: Benefits and problems. *Proceedings of the Sixth International Symposium on Electronic Theses and Dissertations ETD2003*. Berlin: Humboldt-Universität zu.

[27] 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.

[28] Clausen, L.R. 2004. *Handling file formats*. Technical Report. Kongelige Bibliotek.

[29] Christensen, S. 2004. *Archival data format requirements*. Technical Report. Kongelige Bibliotek.

[30] Stanescu, A. 2004. Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *D-Lib Magazine,* 10,11.

[31] Anderson, R., Frost, H., Hoebelheinrich, N., & Johnson, K. 2005. The AIHT at Stanford University. *D-Lib Magazine, 11*(12).

[32] Shirky, C. 2005. *Library of Congress Archive Ingest and Handling Test (AIHT): Final report*. Technical Report. National Digital Information Infrastructure & Preservation Program (NDIIPP).

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

[34] Cornwell Management Consultants. 2005. *Criteria for the selection of preservation formats*. Technical Report. The National Archives, U.K.

[35] Ferreira, M., Baptista, A.A., & Ramalho, J. C. 2006. A foundation for automatic digital preservation. *Ariadne, 48*.

[36] Ferreira, M., Baptista, A. A., & Ramalho, J. C. 2007. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries*, 6, 4, 295-304.

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

[38] Barve, S. 2007. File formats in digital preservation. In Madalli, D.P, & Madalli, P. (Eds.), *International Conference on Semantic Web & Digital Libraries: ICSD-2007*, 239-248.

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

[40] Becker, C., Kulovitz, H. Brown, A. 2008. *Planets: Report on service integration in Plato 2*. Technical Report. Planets Project.

[41] Helmer, O., & Rescher, N. 1959. On the epistemology of the inexact sciences. *Management Science*, 6,1, 25-52.

[42] Dalkey, N.C. 1968. Predicting the future. *National Conference on Fluid Power*, Chicago, Illinois.

[43] Gordon, T. J., & Helmer, O. 1964. *Report on a long-range forecasting study*. Technical Report. RAND Corporation.

[44] Luo, L., & Wildemuth, B. M. 2009. "Delphi studies." In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science*. Westport, CT: Libraries Unlimited.

[45] Delbecq, A.L., Van de Ven, A.H., & Gustafson, D.H. 1975. *Group techniques for program planning*. Glenview, IL: Scott, Foresmann, and Co.