Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

A methodology for measuring the preservation durability of digital formats*

Chao LI†1, Xiao-hui ZHENG[2], Xing MENG[1], Li WANG[1], Chun-xiao XING[1]

([1]*Research Institute of Information Technology, Tsinghua University, Beijing 100084, China*)

([2]*Library of Tsinghua University, Tsinghua University, Beijing 100084, China*)

†E-mail: lichao00@tsinghua.org.cn

Received Sept. 14, 2010; Revision accepted Oct. 10, 2010; Crosschecked Sept. 14, 2010

Abstract: It is now widely recognized that appropriate measures are required for digital preservation to ensure that digital data can be accessed and used currently and in the future. Among all the risks of digital preservation, format obsolescence is one of the most important. There have been several projects or initiatives dealing with the measurement method of format obsolescence risk, but there has been no mechanism to quantify the preservation risk or durability of digital formats based on a self-improving assessment model, executed with the aid of computers. This paper deals with a methodology for measuring the preservation durability of digital formats, especially for their risk assessment. This method is based on a quantitative assessment model for format risk, and can shift the non-quantifiable knowledge or experiences of field experts to a machine identifiable and processible form, or 'risk scores'. Results can be recognized and communicated by computers automatically and formally, which can assist in the automatic/semi-automatic risk management for digital preservation, sharing this quantified knowledge among communities. Because technologies are changing quickly, the quantitative assessment model for risks will not be a status quo situation. Thus, also presented is a method to fine tune the quantitative assessment model for risk of formats through a self-learning and self-improving style.

Key words: Digital preservation, Format obsolescence, Risk assessment model, Risk value
doi:10.1631/jzus.C1001006 **Document code:** A **CLC number:** G250.76; TP319

1 Introduction

Currently, most digital libraries collect massive amounts of digital objects of all types (such as text, image, video, and audio), and store them in diverse hardware with different software for editing and reading. With the rapid development of information technology, these organizations face the problem of how to deal with the obsolescence risks of format, software, and hardware to maintain long-term access to digital collections (Li *et al*., 2008).

To eliminate risks, people directing digital preservation efforts must make decisions and take actions today which will have an impact well into the future. It is safer and cheaper to analyze and compare potential risks and actions before actions are actually taken (Stanescu, 2004).

China-America Digital Academic Library (CADAL) is an international cooperation project directed by computer scientists in China and America. Its aim is to build a digital library with millions resources for education and research (Chen and Zhu, 2005). It can also promote the sharing of global digital resources.

CADAL includes four major parts: digital resource construction, technical support environmental construction, digital library technical center construction, and digitization center construction. As one of the great achievements for the first phase of the CADAL project, one million digital books are provided for users with DjVu format. Also, CADAL

* Project supported by the National High-Tech R & D Program (863) of China (No. 2009AA01Z143) and the Research Foundation of the Ministry of Railways and Tsinghua University, China (No. 20091111068)

chooses TIFF format with G4 compression as a preservation format. Because simplified TIFF format is easy to measure the risk manually, the first phase for the technical support environmental construction has not considered the format risk as a big problem. As one of the missions for the second phase of CADAL, large scale multi-format resources will be processed at several digitization centers in the near three years. Not only do the technical centers need to consider the preservation problem, but also more digitization centers need to think about this for local preservation.

Research has shown that longevity issues for digital resources are defined by two aspects (Stanescu, 2004): (1) the media (e.g., tape, CD-ROM disk) must survive the passage of time; (2) even if media can last a long time, the information must be stored in formats (e.g., JPEG, PDF, TIFF) that can be understood by modern programs. Wotsit.org (http://www.wotsit.org/) lists about 1000 digital formats and versions, many of which are not being used now. Given the speed at which formats emerge and are replaced, how can librarians and archivists identify the preservation risks caused by those formats?

To date, there has been no method to measure the risk of format preservation durability in a quantitative manner. Also, there has been no consensus quantifying the knowledge base of field experts in digital preservation of format risk. Finally, there has been no metrics/scale to communicate or share these measurements with a wider audience in the community. Thus, we need a method to precisely measure the risk of optional formats. As far as we know, there has been no mechanism to quantify the preservation risk or durability of digital formats based on a self-improving assessment model that can be executed with the aid of computers.

In this paper, we propose a methodology for measuring the preservation durability of digital formats. This method is based on a quantitative assessment model for format risks, and can shift the non-quantifiable knowledge or experiences of field experts to machine identifiable forms of ‘risk scores’. This kind of results can be recognized and communicated by computers automatically and formally, which can help in the automatic/semi-automatic risk management for digital preservation, and the easy sharing of this quantified knowledge among communities.

As technologies are evolving rapidly, the quantitative assessment model for risks cannot be set in stone. This paper also presents a way to fine tune the quantitative assessment model for format risks through a self-learning and self-improving style.

2 Related works

For the long-term preservation of digital resources, many libraries, archives, and other related research organizations abroad have researched extensively both in theory and practice. They have created many related projects, trying to research different parts of long-term preservation. Some projects are done to create technical standard specifications, such as PRONOM (http://www.nationalarchives.gov.uk/pronom/) and Journal Storage/Harvard Object Validation Environment (JHOVE), others to research the metadata, such as the Preservation Metadata Implementation Strategies (PREMIS). Some projects deal with intellectual property rights, such as the Rights Metadata for Open Archiving (RoMEO), and others with storage systems, such as an Approach to Digital Archiving and Preservation Technology (ADAPT). Finally, some projects concentrate on preservation warehousing, such as Archival Digital Libraries Repositories, and also planning at the international level, such as the National Digital Information Infrastructure and Preservation Program (NDIIPP), the Joint Information Systems Committee (JISC), and the Electronic Resource Preservation and Access Network (ERPANET) (Automated Obsolescence Notification System, AONS, http://www.apsr.edu.au/aons).

The National Library of Australia is developing mechanisms specifically focused on monitoring and assessing the risks of file format obsolescence. They made the AONS II project (Zhao, 2004) in conjunction with the Australian Partnership for Sustainable Repositories (APSR). The project was aimed to develop a software tool that allows users to automatically monitor the status of file formats in their repositories, make risk assessments based on a core set of obsolescence risk questions, and receive notifications when file format risks change or other related events occur. AONS II made a model to evaluate format obsolescence risks, but the system has not been wholly achieved.

Besides, Cornell University made the Virtual Remote Control (VRC) project (Kenney *et al*., 2002; http://irisresearch.library.cornell.edu/VRC/) based on a list covering several aspects that may cause risks. VRC attempts to discover how specific Web resources, documents, and Web sites change over time, in order to anticipate their potential disappearance (Stanescu, 2004). In the risk measuring process, VRC added the participation of experts. The process includes risk definition, risk classification, risk measuring, risk analysis, and risk management (Li *et al*., 2008; 2009). Experts in the system should participate in definition, classification, and analysis processes. Though the model of VRC listed several factors that may affect the measuring process, it was not changed into a language that could be realized by a computer. The National Science Library of the Chinese Academy of Sciences developed a long-term preservation system, based on Fedora, to preserve digital resources. The system followed the Open Archival Information System (OAIS) model, and created an integrated frame to manage the lifecycle of digital object preservation. As a result, the system, including the import module, preservation management module, and access module, could preserve many different kinds of digital resources flexibly.

3 Description of the format risk assessment system

This format obsolescence risk assessment system includes three subsystems, namely the object recognition subsystem, risk assessment questionnaire subsystem, and risk assessment model for optimization subsystem (Fig. 1). These three subsystems test the existence of obsolescence risks of digital formats at different stages. The main functions of the subsystems are: identifying and obtaining digital format information by scanning the document in the file storage system; making online surveys of experts in the field and librarians according to two sets of risk assessment questionnaires (community-risk and local-risk parts, the former for domain experts, the latter for local librarians); quantifying the results of risk assessment; and, assessing format obsolescence risk by scoring the risk. The system, meanwhile, recognizes the functions of a series of user managements according to online surveys and background management of questionnaire answers.

This paper describes the method used in the third subsystem—a risk assessment model for optimization subsystem. In addition to realizing Web applications of changes to questionnaires options, the system mainly develops an obsolescence risk assessment algorithm of digital format.

Preliminary work presents the two questionnaires (community-risk and local-risk parts) as a Web page to readers. Prior to entering the risk assessment model for optimization subsystem, format information (e.g., format version) has already been collected, completed by the former two subsystems. Through this system, risk assessment results will be given as the value of obsolescence risk for digital formats. The questionnaire comprises several subjective subjects on factors that constrain the evaluation of obsolescence risk. This stage is used to initialize and optimize options, and then to calculate the value of

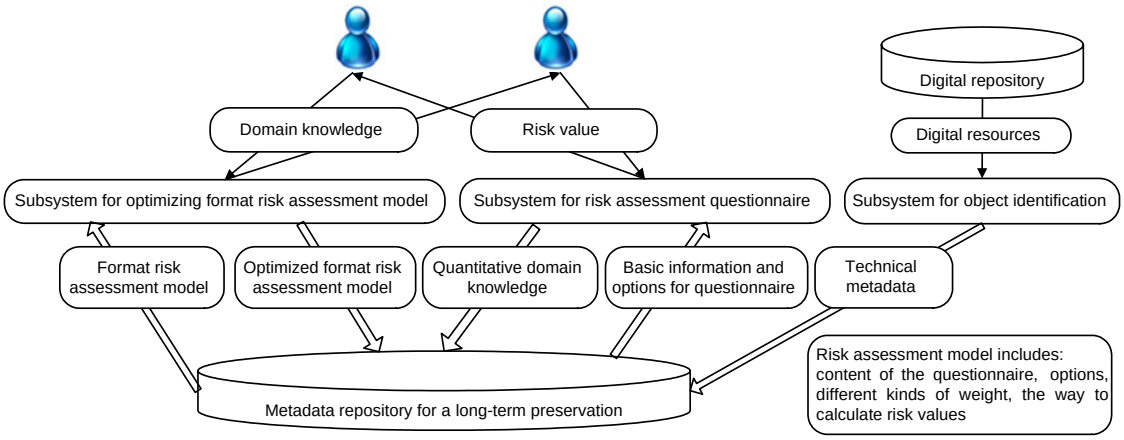

Fig. 1 Format risk assessment system

obsolescence risk according to options. The role of optimizing options is to reduce the deviation resulting from inappropriate design for the options. In addition, as time goes by, options need to adjust to actual situations because the risk assessment factors may change. We designed the format obsolescence risk evaluation module into Web forms. Questionnaires are presented to readers dynamically in the form of Web pages, and the achievement of a risk value algorithm is one part of it. Users enter the evaluation system through some sort of identification (experts in the field or librarians, etc.). Different identities possess different permissions, with the administrator possessing the highest privileges. The obsolescence risk value of a certain format is obtained by searching and assessing the format in a range (such as a folder).

In this subsystem of the evaluation algorithm, administrators are permitted to originally change subject options according to the answers of experts. After the options of subjects in the questionnaires are determined, a set of model answers is rendered and the risk value of a format is calculated based upon this, using the clustering algorithm in data mining by different experts in the field. It is important to note that an average risk value is obtained either by taking each risk value as the final certain risk value, or by making a decision to obtain a relatively standard reference answer. A risk value is then calculated via the above algorithm as the final risk value. We still need more tests to find out if results can be more convincing in practical applications.

4 Description of the method

The methodology includes three steps: (1) initialize the questionnaire; (2) calculate and determine the model parameters according to the questionnaire, structure and optimize the measuring model and synthesize answers of the experts; and, (3) calculate the value of the risk of format obsolescence for certain formats.

4.1 Step 1: initialize the questionnaire

According to the features of long-term conservation domain and the domain knowledge in this field, the system designers define the influencing factors to measure the format obsolescence risk value. The knowledge is included in the content of the questionnaire. Then we set the number N and the initial status of the options for each question. The initial status is equal to the initial threshold. The precondition on which this method depends is that the number N and the initial status of the options of each question are defined by the experienced designers, who have a certain amount of domain knowledge. Thus, the number of the options will not be too large or the initial status of the options deviates from the actual conditions.

4.2 Step 2: establish the assessment model

In this step, we fix and calculate the model parameters, establish the assessment model, and optimize it.

The process can be divided into three stages.

Stage 1: Define the effect weight of every question in the questionnaire on the overall format obsolescence risk value.

Every question's weight of effect on the overall format obsolescence risk value is not identical; some questions have more fatal effects, while others have less. Hence, we need a method to define every question's weight of effect on the overall format obsolescence risk value.

This method uses the classical algorithm, ReliefFAttributeEval (Robnik-Šikonja and Kononenko, 1997), to define each question's influential weight on the format obsolescence risk in the questionnaire. In ReliefFAttributeEval, the answer to each question that all experts in this field give, and the overall estimate of the format obsolescence risk of a format, are both regarded as samples of the data concentration. A number of samples constitute a sample set. Each question in the questionnaire corresponds to one attribute of ReliefFAttributeEval. The overall judgment of the format obsolescence risk of a given format corresponds to the objective class of ReliefFAttributeEval. The sample set is used to determine the influential weight of the attributes to the objective class. Hence, when using ReliefFAttributeEval, the experts in this field need to give an overall estimate of the format obsolescence risk before or after experts in this field answer the assessment questionnaire (the risk is considered low if the overall judgment is in the interval $(x, (x+y)/2]$, and high if the overall judgment is in the interval $((x+y)/2, y]$). Then ReliefFAttributeEval is used to determine each question's influential weight on the overall format obsolescence risk value.

This stage has four features:

1. No need to limit the data types of attributes.

2. Not sensitive to the relationship between attributes. Among the questions of the questionnaire, there is a certain dependence, which corresponds to the relationship of attributes in ReliefFAttributeEval. Many methods that presume the attributes are independent of each other do not fit the applicative situation of the method.

3. No need to eliminate the redundant attributes. ReliefFAttributeEval gives a certain influential weight to all the attributes related to the objective class, even if the attributes are redundant with another one. This feature fits perfectly the applicative situation of this method, because it is supposed that each question in the questionnaire has a certain influence on the format risk, and that the questionnaire is not a set of useless and irrelevant questions.

4. High running efficiency. The computation complexity of ReliefFAttributeEval is $O(tmN)$, where t is the number of trials, m is the number of samples, and N is the number of back-up attributes. Compared with many other methods, the amount of calculations is quite small.

Stage 2: Determine the threshold of each option for each question.

In this method here, the influential weight value of each option in the questionnaire is equal in different distributions. It will influence the quantitative accuracy of the overall format risk assessment model if the range and precision of each option's threshold are not adapted to the actual technological situation. Hence, the threshold of each option needs to be adjusted as the technology is updated.

This stage has three parts:

1. According to the domain knowledge and experience, the system designer determines the number N (a positive integer) and the initial status of the options for each question. The initial status means the initial threshold or the subjective judgment. A precondition on which this method depends is that the number N and the initial status of the options of each question are defined by the experienced designers who have a certain amount of domain knowledge. Thus, the number of the options will not be too large or the initial status of the options deviates from the actual conditions.

2. If there are K experts joining the questionnaire investigation of a given format, and more than $K/2$ experts choose a same option P_i (i is the order number of the question's options, i=1, 2, …, N), it indicates that the current option's thresholds are not exact enough. That is, the partitioning granularity is too large, and the option partition does not attain the effect that the disparity of the risk values of the options of a question is distinguished. The threshold of the option needs to be reduced to P_i-d ($0\le d\le P_i$). In addition, other option thresholds are adjusted slightly to balance the probability of choosing each option, and to reduce the influence on the format obsolescence risk value as the options' thresholds are divided improperly.

3. To slightly adjust the option thresholds according to the distribution situation of the questionnaire answers given by these experts, after adjusting, the content of the questionnaire is updated, and the experts need to answer the updated questionnaire again. When the option thresholds of the updated questionnaire become rational, i.e., the conditions in Step 2 do not occur, the task of adjusting is accomplished; otherwise, the option threshold is again adjusted slightly following Step 3.

Stage 3: Define the format obsolescence risk value for each option of every question.

The different options for the same question represent different levels of obsolescence risk influence severity. Thus, we need to distinguish the differences for these distinct options by defining their risk values.

This stage can be divided into two parts:

1. For a given question E in this questionnaire, whose N options are shown in turn: $P_1...P_i\ldots P_n$ ($1\le i\le N$), sort these N options in the order of their format obsolescence risk severity.

2. Calculate these format obsolescence risk values of N options sorted. The method is shown below. The option P_i''s ($1\le i\le N$) format obsolescence risk value is

$$V_{P_i'}=\frac{W_E}{\sum_1^M W_i}\cdot(y-x)\cdot\frac{i}{N},$$

where W_E is the influence weight of the format obsolescence risk of P_i''s question E, which is calculated in Stage 1, $\sum_1^M W_i$ is the sum of the format

obsolescence risk influential weights of all questions in the questionnaire, M is the number of the questions in the questionnaire, and $y-x$ is the format obsolescence risk threshold span to which the format obsolescence risk quantification interval $(x, y]$ corresponds.

4.3 Step 3: calculate the format obsolescence risk value

In this step, we synthesize answers of the questionnaire the experts gave, and then calculate the format obsolescence risk value according to the assessment model.

For each format, this method needs five or more experts in this field to answer the questionnaire to ensure the reliability and operability of the format obsolescence risk value obtained. Therefore, we need to synthesize those questionnaire answers given by the experts in the fields to obtain the final format obsolescence risk value.

There are two methods to obtain the final format obsolescence risk value.

The first method: calculate the format obsolescence risk value according to the answers given by the experts in this field, and then synthesize them to obtain the final format obsolescence risk value.

1. Calculate the format obsolescence risk values of the experts for a particular format. The experts calculate the format obsolescence risk value by

$$\text{risk} = y - \sum_{1}^{M} V_t,$$

where y is the upper limit of the format obsolescence risk value quantization interval $[x, y]$, and V_t is the format obsolescence risk value of the answer that a given expert provides to a given question, $\exists V_t$, $V_t \in \{V_{P'_{t_1}}, \ldots, V_{P'_{t_i}}, \ldots, V_{P'_{t_N}}\}$ $(1 \le t \le M, 1 \le i \le N)$.

2. Take the obsolescence risk values that the experts work out for the format as a sample of the SimpleKmeans method, a classic algorithm in data analysis, and a sample set consisting of a few samples. Use the SimpleKmeans method to obtain the cluster center of the sample set, i.e., the final format obsolescence risk value.

The second method: synthesize the answers given by the experts in this field, and then obtain the standard answers to calculate the final format obsolescence risk value.

1. For each question in the questionnaire, take the answer any expert gives as a sample, and the sample set consists of the answers the experts give to the same question. Use the SimpleKmeans method to obtain the cluster of the sample set, and determine the standard answer for every question. Then obtain the standard answer for the whole questionnaire.

2. According to the standard answer for the whole questionnaire, calculate the final obsolescence risk value for the format. According to the standard answer of the questionnaire, calculate the format obsolescence risk value by $\text{risk} = y - \sum_{1}^{M} V_t$, where y is the upper limit of the format obsolescence risk value quantization interval $[x, y]$, and V_t is the format obsolescence risk value of the standard answer of some question in the questionnaire, $\exists V_t, V_t \in \{V_{P'_{t_1}} \ldots V_{P'_{t_i}} \ldots V_{P'_{t_N}}\}$ $(1 \le t \le M, t_1 \le t_i \le t_N, t_1=1$, and t_N is the number of the options of question t).

The features of the second method are listed below:

1. Excellent effect. The SimpleKmeans method tries to find the K distributions that produce the lowest squared error function. If the target class is intensive, its effect is good. For our situation, the result of the experts in this field will not be extraordinary; thus, we consider the result class as intensive.

2. Simple, fast, and effective. The complexity of the SimpleKmeans method is $O(nkt)$, where n is the number of objects (in this case, the number of experts in this field who complete the questionnaire).

5 Implementation of the method

5.1 Step 1: initialize the questionnaire

In the system we designed, we confine the format obsolescence risk value range to the interval (0, 10]. If risk<1 or risk>10, the questionnaire system will exit. Before entering the questionnaire system, the initiator will give the format to be evaluated (denoted as F) an initial risk value (denoted as y) 10. The value of y will be reduced according to the options the experts choose. In our prototype system, we use the questionnaire of AONS (Pearson, 2008), and modify it by adding the choices A–D to the original open questions. The results of initializing the questionnaire are shown below:

Q1: Is this a base format or an ubiquitous format which is likely to be rendered by most applications?
A. Yes B. No
Q2: Is this file format and version referenced in any searched information resources?
A. Yes B. No C. I have no idea
Q3: Is there a known support end date for this format version?
A. Less than 5 years B. Between 5 and 10 years
C. More than 10 years D. I have no idea
Q4: How many years since this version was released?
A. Less than 5 years B. Between 5 and 10 years
C. More than 10 years D. I have no idea
Q5. How many new versions have been released since then?
A. Not more than 2 B. 3 or 4
C. 5 or more D. This format is abolished
Q6: Overall, how many access options are effectively available to it, including the original rendering software?
A. Only one B. 2 or 3
C. 4 or more D. I have no idea
Q7: Are there critical hardware and software dependencies for effective use of the original rendering software?
A. No requirements in software and hardware
B. Loose requirements in software and hardware
C. Strict requirements in software and hardware
D. I have no idea about the requirements in software and hardware
Q8: How many alternative software options for safe and effective rendering can be identified?
A. Not more than 2 B. More than 2 C. 0 or unknown
Q9: For each alternative, are there critical hardware and software dependencies for effective use of the alternative rendering software?
A. Most back-up tools have no requirements in software and hardware
B. Most back-up tools have loose requirements in software and hardware
C. Most back-up tools have strict requirements in software and hardware
D. I have no idea

5.2 Step 2: establish the assessment model

1. Determine the influential weight of the nine questions. Before answering the assessment questionnaire, every expert needs to give an overall estimate of the obsolescence risk of the format *F*. The risk is considered low if the overall judgment is in the interval (0, 5], and high if the overall judgment is in the interval (6, 10]. Using ReliefFAttributeEval, one of data concentration methods, the answers to the nine questions from all experts in this field, and the overall estimate of the format obsolescence risk of a given format by all experts in this field, are regarded as a sample of the data concentration. The nine questions correspond to the attributes of ReliefFAttributeEval. The overall judgment of the format obsolescence risk of a given format corresponds to the objective class of ReliefFAttributeEval. When using ReliefFAttributeEval, each parameter is the default value, and then one obtains the influential weight of each question on the format obsolescence risk value, i.e., W_i, $i\in[1, 9]$.

2. To define the threshold of the different options of these nine questions, it is essential to pass these steps: (1) Experts in this field answer the initialized questionnaire model. (2) Estimate if the questionnaire model achieves the optimization goal according to the situation of the answers of the questionnaire. For a given format, *K* experts combine to answer the questionnaires, the questionnaire do not achieve the optimization goal if more than *K*/2 experts answers focus on some same options. (3) Then slightly adjust this option range, i.e., narrow the range of this option to achieve the goal of optimizing model. (4) After optimizing the questionnaire model, experts in this field need to answer the optimized questionnaires again, until the model achieves the optimization goal; i.e., when the number of the chosen times for every option is less than half the number of experts in this field who have answered the questionnaires, the optimization work is accomplished; otherwise, this process should be circulated, until the optimization goal is achieved.

3. Determine the influential weight of the format obsolescence risk value for each option of the nine questions. After the options are adjusted slightly and determined, sort the options for each question in descending order of the influence of the format obsolescence risk value. The risk value of each option sorted is then determined. Taking question *E* as an example, the sorted options are $P_1 \ldots P_i \ldots P_N$, where the format obsolescence risk value of the option is

$$V_{P_i} = \frac{W_E}{\sum_1^9 W_i} \times 10 \times \frac{1}{4}.$$

Herein W_E is the influential weight of question *E*, which is obtained by the method described in Section 4.2; the number of the questions is M=9; 10 is the format obsolescence risk threshold span to which the format obsolescence risk value quantitative interval (0, 10] corresponds; and N=4 in the general situation (i.e., the options are ABCD, four options). Hence, the reduced risk value amount is V_{P_i} when choosing the option for question *E*. The calculation model of the format obsolescence risk value of the questionnaire in this example is constructed accordingly (Fig. 2).

5.3 Step 3: calculate the format obsolescence risk value

In this step, there are two methods to calculate the format obsolescence risk value.

5.3.1 First method

1. Calculate the format obsolescence risk value the experts in this field gave the format *F*. Obtain the format obsolescence risk value the experts in this field gave the format *F* according to Steps 1 and 2. The set of the format obsolescence risk values which the experts gave to the format *F* is {$risk_1$, $risk_2$, …, $risk_K$, ...}, where $risk_K$ is the format obsolescence risk value the *K*th expert give the format *F*, which is obtained in Step 2.

2. Take the obsolescence risk values {$risk_1$, $risk_2$, …, $risk_K$, ...} that the experts work out for the format *F* as a sample of the SimpleKmeans method,

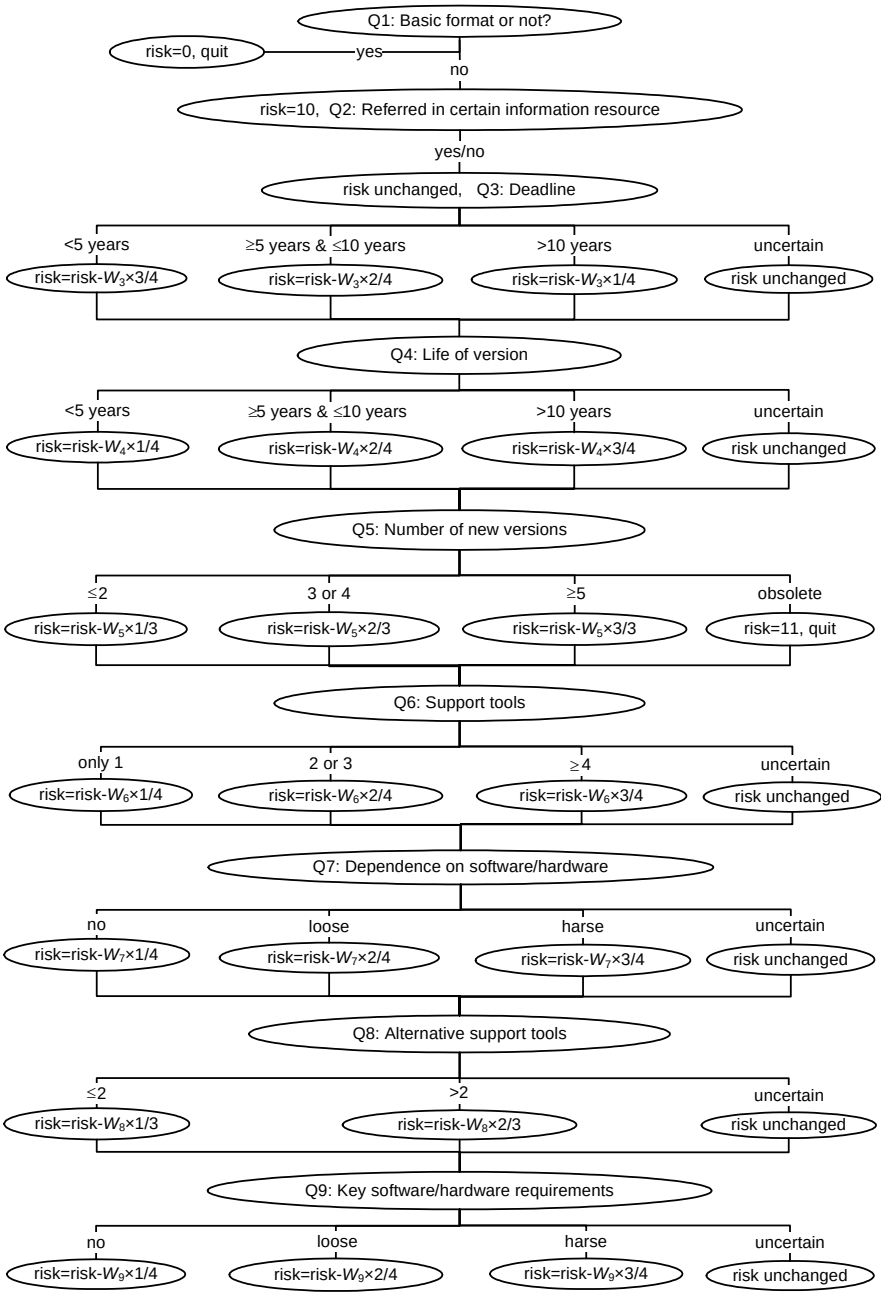

Fig. 2 Quantitative model for community risk

and a sample set consisting of a few samples. Use the SimpleKmeans method to obtain the cluster center of the sample set, i.e., the final format obsolescence risk value.

5.3.2 Second method

1. For each question in the questionnaire, take the answers that experts give to the same question as a sample set of the SimpleKmeans method. Use the SimpleKmeans method to obtain the cluster center of the sample set, i.e., the standard answer of each question. Following this way, the standard answer set of the whole questionnaire $\{F_1, F_2, ..., F_9\}$ for the format F can be obtained.

2. According to the standard answer of the entire implementing example questionnaire, calculate the final format obsolescence risk value of the format F following the above Steps 1 and 2 described in Section 4, i.e., the format F's final format obsolescence risk value:

$$\text{risk} = 10 - \sum_{i=1}^{9} V_{F_i},$$

where V_{F_i} is the format obsolescence risk value of the ith question in the standard set of the format F, and is obtained in Step 2, and 10 is the initial format obsolescence risk value of the format F.

6 Test results

We invited the experts of Tsinghua University Library to test our method. They evaluated the format obsolescence risk value of seven formats—ppt, text configuration file (ini), JPEG file interchange format (jfif), Windows shortcut file (lnk), extensible hypertext markup language (xhtml), hypertext markup language (html), and OS/2 Bitmap (bmp). We just show the result of format lnk to represent the results of the evaluation (Fig. 3).

As the values in Fig. 3 show, we conclude that the final risk value is 8.0 (the first method of the algorithm) or 6.0 (the second method of the algorithm). In our opinion, 6.0 is better to show the true risk of lnk.

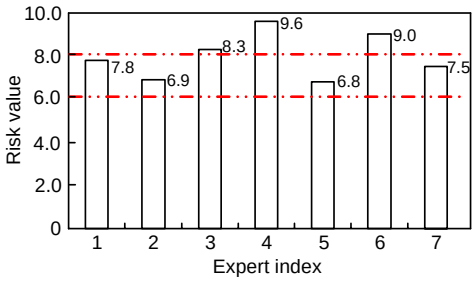

Fig. 3 Risk value of format lnk

7 Conclusions

We propose a methodology for measuring the preservation durability of digital formats, focusing on the risk assessment. The method is based on a quantitative assessment model for risk of formats, and shifts the non-quantifiable knowledge or experiences from field experts to a machine identifiable and processible form or 'risk scores'. Results can be recognized and communicated by computers automatically and formally, which can assist in the automatic/semi-automatic risk management for digital preservation and sharing quantified knowledge among the communities easily.

As technologies are changing quickly, the quantitative assessment model for the risks must evolve over time. This paper also presents a way to tune the quantitative assessment model through a self-learning and self-improving style. Based on the model and the tuning method, we designed and implemented a Web-based application for the format risk management.

In conclusion, we believe that the final implementation of the system will be a significant advance towards the format risk management for digital preservation. This method, based on a quantitative assessment model and self-improving tuning, will effectively assist in making preservation decisions.

8 Future directions

Obviously, the process of measuring the preservation durability risk of digital formats is best done in collaboration with all the related parties, e.g., holders of repositories, digital libraries, digital

archives, global format/software registries, format/software/hardware experts, and researchers. We hope that all interested parties will contribute through our system.

Over time, with more and more data collected in the system, this method and its application can provide more precise results, and thus lead to more reliable decisions on digital preservation. These cumulated data and corresponding results should be kept in a publicly available platform (maybe with format and software registries) to allow widespread contribution and sharing.

Lastly, with the passage of time, the history of these data and corresponding results could be used to predict and analyze technology trends, and to allow the participants to make decisions and take actions as late or as early as necessary.

References

Chen, H.Y., Zhu, H.K., 2005. A summarization on the China-America Digital Academic Library. *J. Acad. Libr.*, (1):3-6 (in Chinese).

Kenney, A.R., McGovern, N.Y., Botticelli, P., Entlich, R., Lagoze, C., Payette, S., 2002. Preservation risk management for Web resources. *D-Lib Mag*. [doi:10.1045/january2002-kenney]

Li, C., Ma, N.N., Xing, C.X., Jiang, A.R., 2008. An integrated approach for smart digital preservation system based on Web service. *LNCS*, **5362**:347-350. [doi:10.1007/978-3-540-89533-6_41]

Li, C., Xing, C.X., Dong, L., Huang, M.B., 2009. A Semi-Automatic System for Managing Multiple Digital Preservation Risks of Digital Libraries in China. Proc. 9th ACM/IEEE-CS Joint Conf. on Digital Libraries, p.425. [doi:10.1145/1555400.1555495]

Pearson, D., 2008. AONS II: continuing the trend towards preservation software 'Nirvana'. *New Technol. Libr. Inf. Serv.*, (1):42-49.

Robnik-Šikonja, M., Kononenko, I., 1997. An Adaptation of Relief for Attribute Estimation in Regression. Proc. Int. Conf. on Machine Learning, p.296-304.

Stanescu, A., 2004. Assessing the durability of formats in a digital preservation environment: the INFORM methodology. *D-Lib Mag*. [doi:10.1045/november2004-stanescu]

Zhao, J.L., 2004. Australia network information preservation project PANDORA and its enlightenment. *Inf. Stud. Theory Appl.*, **27**(5):552-554 (in Chinese).