International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:13, No:11, 2019

A Bayesian Classification System for Facilitating an Institutional Risk Profile Definition

Roman Graf, Sergiu Gordea, Heather M. Ryan

Abstract—This paper presents an approach for easy creation and classification of institutional risk profiles supporting endangerment analysis of file formats. The main contribution of this work is the employment of data mining techniques to support set up of the most important risk factors. Subsequently, risk profiles employ risk factors classifier and associated configurations to support digital preservation experts with a semi-automatic estimation of endangerment group for file format risk profiles. Our goal is to make use of an expert knowledge base, accuired through a digital preservation survey in order to detect preservation risks for a particular institution. Another contribution is support for visualisation of risk factors for a requried dimension for analysis. Using the naive Bayes method, the decision support system recommends to an expert the matching risk profile group for the previously selected institutional risk profile. The proposed methods improve the visibility of risk factor values and the quality of a digital preservation process. The presented approach is designed to facilitate decision making for the preservation of digital content in libraries and archives using domain expert knowledge and values of file format risk profiles. To facilitate decision-making, the aggregated information about the risk factors is presented as a multidimensional vector. The goal is to visualise particular dimensions of this vector for analysis by an expert and to define its profile group. The sample risk profile calculation and the visualisation of some risk factor dimensions is presented in the evaluation section.

Keywords—linked open data, information integration, digital libraries, data mining.

I. Introduction

IN recent years, libraries, archives and museums have created new digital collections that comprise millions of objects, with the goal of making them available on a long term basis. One of the core preservation activities deals with the evaluation of appropriate formats used for encoding digital content. The preservation risks for a particular file format are often difficult to estimate as concluded in [6]. The definition of risk factors and associated metrics is still an open research topic in the digital preservation community. Involvement of digital preservation experts is required for collecting complete information and evaluating preservation risks [1]. Currently, individual institutions select their own file formats for long term preservation depending on particular projects, preservation goals, workflows and assets. Due to the scale of digital information that has to be managed, memory institutions are facing challenges regarding preservation, maintenance, and quality assurance of these collections. For that reason, automated solutions for data management and digital preservation are absolutely necessary. Many file formats are properly documented, are open-source and well supported by software vendors. However, other formats may be outdated or no longer functional with modern software or hardware. There are also custom/proprietary formats - which may be obsolete and not renderable with commodity hardware. To address these problems, we employ the File Format Metadata Aggregator (FFMA) system [5] and an information integration approach. Therefore, the risk factor estimation and settings is an important open issue. The proposed approach facilitates the definition of institutional risk profiles. The institutional risk profile can be defined as decisions made by decision-maker within an institutional context. The goal of this approach is to help build an institutional risk profile. The novelty of this technical solution is the employment of data mining methods to facilitate complex risk factor settings for preservation experts. These methods support edangerment group evaluation for the inititutional risk profile based on the Bayes theorem. Decision support based on the elaborated rule engine provided by FFMA, fuzzy rules and an expert knowledge base is designed to support institutions like libraries and archives with assessment for analyzing their digital assets. The factors for the risk metrics and classifier calculation were provided through a study organised by Heather Ryan [14]. This paper is structured as follows: Section II gives an overview of related work and concepts. Section III explains the risk factor visualisation workflow and also covers data mining issues. Section IV presents the experimental setup, applied methods and results. Section V concludes the paper and provides an outlook on planned future work.

Roman Graf and Sergiu Gordea are with the Department Digital Safety & Security, Austrian Institute of Technology GmbH, Austria, (e-mail: {roman.graf,sergiu.gordea}@ait.ac.at).

Heather M. Ryan is with Library & Information Science Program, University of Denver (e-mail: heather.m.ryan@du.edu).

II. Related Work

The research on risk management in digital collections increasingly gains in importance. It is difficult to guarantee the longivity of digital information. The investigation [12] aims at risk assessment of migrating of file formats. Accurate format identification and rendering is a challenging task due to malformed MIME types, rendering expenses, dependence on content not embedded in the file, missing colour tables, changed fonts, etc. In [10], the author examines how the network effects could stabilise formats against obsolescence. The result of evaluation demonstrates that most formats last much longer than five years, that network effects stabilize formats, and that new formats appear at a modest, manageable rate. However, a number of formats are fading from use and every corpus contains its own biases. Digital preservation tools

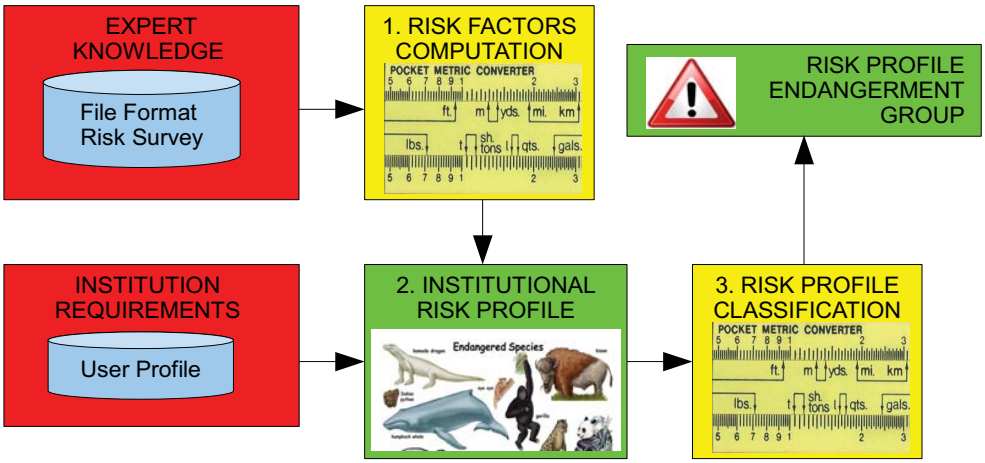

Fig. 1: The overall workflow for the endangerment analysis of the institutional risk profile.

like PANIC [9], AONS II [13], SPOT [16], P2 registry [2], aimed at identifying file formats used for encoding digital collections and informing repository managers of events that might impact access to the stored content. They also define alerting mechanisms when file formats become obsolete. As distinct from our approach they do not apply expert knowledge and do not specify risk factors that may influence file format endangerment. The FFMA [4] is a preservation planning tool that offers assessment for long-term preservation of digital content. This tool performs an analysis of file formats based on the concept of risk scores. Selected institutional risk profile in conjunction with FFMA can calculate endangerment risks for selected file format. There are multiple influential algorithms [17] (k-Means, SVM, kNN), which can be applyied in data mining. The Naive Bayes is one of them and it is very good matching for classification task in our approach. Bayesian networks [8] extended with statistical techniques are used in data mining to encode probabalistic relationships among variables of interest. Such networks combine prior uncertain expert knowledge with the data and are related to graphical modelling techniques for supervised and unsupervised learning and for learning with incomplete data. In our approach we do not use rule bases, decision trees or artificial neural networks but employ the Naive Bayes method for probabilities calculation. In the proposed approach we intend to apply standard statistics and data mining methods for digital preservation tasks. The proposed system is unique for the given domain.

III. Institutional Risk Profile Definition System

Fig. 1 shows the general workflow for the endangerment analysis of an institutional risk profile.

A. Risk Factor Analysis and Visualisation

The data for risk factor calculation (step 1 in Fig. 1) is aggregated from the expert knowledge base. The knowledge base employs the analyzed risk factors and the institutional requirements with user profile for the institutional risk profile

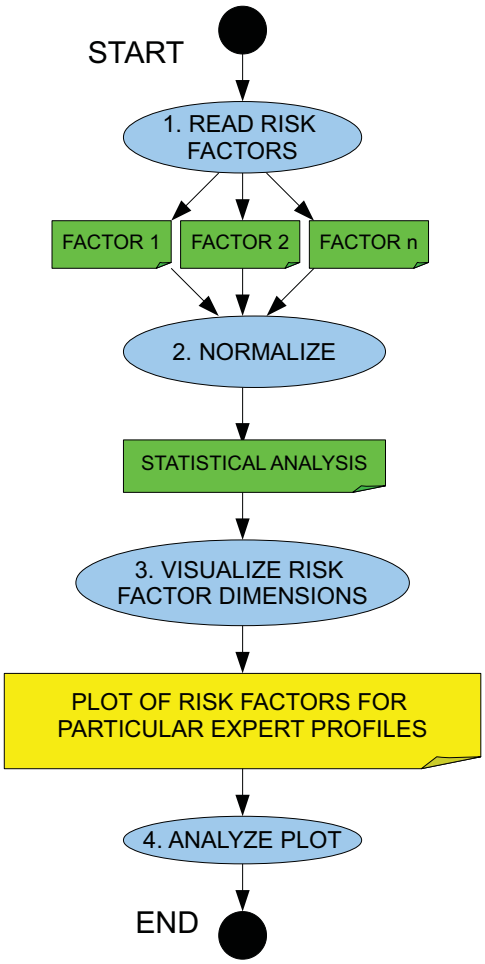

Fig. 2: The risk factor visualization workflow.

computation. Classification of the institutional risk profile based on Bayes theorem provides estimation about the endangerment group of the risk profile. Each risk profile is represented by a multidimensional vector. In the presented approach 28 dimensions are aggregated by the domain experts. The risk profile visualisation is conducted according the workow shown in Fig. 2.

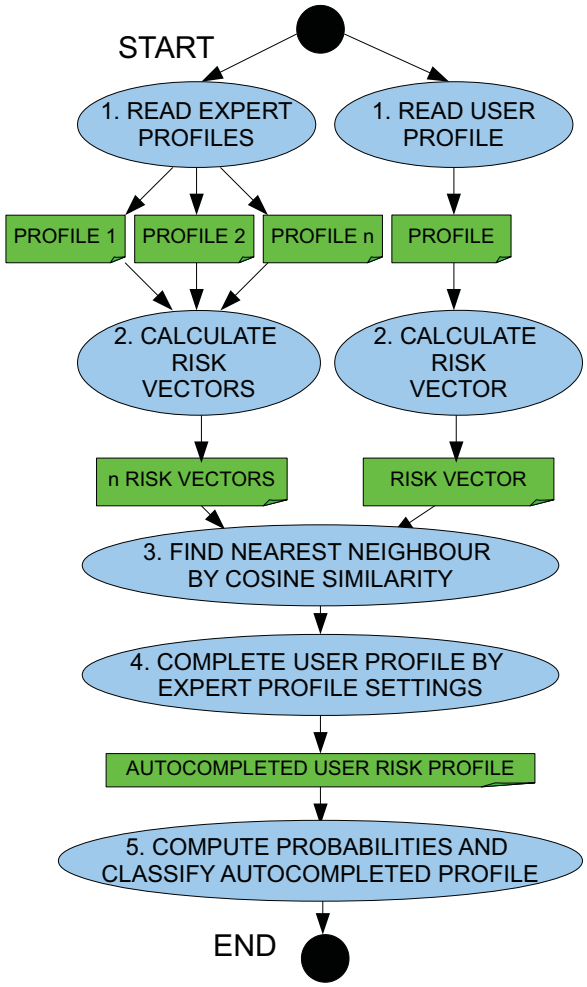

Fig. 3: The workflow for recommendation of a user risk profile.

The risk profile data acquired from domain experts is structured and stored for reuse in computations. The workflow execution starts with the reading of this data from the file. The workflow employs a data mining method that calculates distances between risk profiles based on the values of their risk factors. The scale of the different risk factors is different, e.g. a risk factor can be measured in range 1 to 3 or 1 to 5 or 1 to 100 or boolean value. Therefore, in order to get rid of a mismatch of scale between the features and have a possibly well-balanced risk factor rf set, the workflow applies normalisation in the second step.

$$MSS_i = \frac{rf_i - \mu}{\frac{1}{card(rf)} \sum |rf_i - \mu|}. \quad (1)$$

Normalisation employs the modified standard score (MSS) [15], [19] (see 1), which prevents the influence of the outliers. Each risk profile column is normalised separately. The modified standard score demonstrates how big the deviation from the median value μ is. First the median value for each column is calculated. The median is a middle value from the list, arranged from lowest to highest value. Then, based on the median, the absolute standard deviation can be calculated. In the third step computed risk factors are visualised for the given dimension. E.g. one such dimension is a relation between expert 3 and expert 5 risk profiles. Finally a human expert should analyse the resulting plot in the context of a particular preservation planning task.

B. Risk Profile Recommender

The calculation of the nearest risk profile (step 2 in the workflow on Fig. 1) is described in the workow shown in Fig. 3. The risk profile data manually aggregated by domain experts is stored in a text file and is used in the classification task. Contrary to the visualisation case, each column depicts a particular risk factor. And rows in this file comprise values for the associated risk profile. The institutional risk profile that comprises the most important factors for the institution settings is stored in a text file. The workflow execution starts with the reading of input files. Input risk profiles are stored in the data model and are converted in risk profile vectors in the second step.

$$cos(x, y) = \frac{(\sum x_i \cdot y_i)}{\sqrt{(\sum {x_i}^2 \cdot \sum {y_i}^2)}}. \quad (2)$$

Applying the cosine similarity algorithm (see 2 [11], [3], [18] we find the nearest risk profile from the expert knowledge base. In the next step we merge the detected nearest risk profile with institutional settings and produce the autocompleted institutional risk profile.

C. Risk Profile Classification

Having the autocompleted institutional risk profile at hand we can classify (step 3 in the workflow on Fig. 1) this profile to one of the risk profile groups employing the naive Bayes algorithm [20] (see 3). This formula shows the probability of the risk profile R (4) being the risk profile group g. The risk profile R is a product of all risk factors rf that are comprised in the associated risk profile.

$$p\left(g|R\right) = \frac{p\left(R|g\right)p\left(g\right)}{p\left(R\right)}. \quad (3)$$

$$R = \left(rf_1, rf_2 ... rf_{28}\right). \quad (4)$$

The naive Bayes algorithm picks the risk profile group with the highest probability. The possible hypotheses for risk profile group calculation are that an institutional risk profile has one of the three levels “HIGH”, “MIDDLE” or “LOW”.

We would like to know whether the selected insitutional risk profile belongs to the "HIGH", "MIDDLE" or "LOW" risk profile group. In order to compute the probability of the particular hypothesis given associated risk factors, we multiply the individual probabilities.

Therefore, an expert who's risk factor settings match a particular expert risk profile is more likely to have an associated risk profile group for estimation of the further risk factor settings in his decisions for the digital preservation long-term planning. There are three risk profile groups for institutional expert risk profiles. The type "HIGH" indicates that the expert considers the overall file format endangerment level as very high. The "MIDDLE" risk profile type determines the influence of the risk factors as moderate. And finally, the "LOW" risk profile type means that an expert in average believes that mentioned risk factors does not have significant impact on the endangerment level. The risk profile types are in this case categorical data not numerical.

The risk profile group is evaluated automatically by summarising all risk factor ratings for each expert and splitting the expert profiles in three groups by the resulting risk factor value. For this classification, besides the risk factor ratings additional data can be added and additional types can be defined.

The output of the classifier training should contain a list of prior and conditional probabilities. In our approach the prior probabilities are P(high)=0.2, P(middle)=0.4 and P(low)=0.4. For the Bayesian approach we assume that risk factors are independent.

IV. Evaluation

The goal of this evaluation was the leveraging of the domain expert knowledge base for detection of the nearest risk profile as described in the workflow for autocompletion of a user risk profile (see Fig. 3) and exploitation of aggregated data for visualisation of risk factor coherences. This process is described in the risk factor visualization workflow (see Fig. 2).

A. Hypothesis and Evaluation Methods of the Risk Factor Analysis

The hypothesis is that similar risk factor profiles automatically aggregated from a domain expert knowledge base are located close to each other in the plot for a particular dimension. Therefore, a human expert can easily detect alternative risk factor profiles with particular features for a specific task. Our approach should give an organisation a base of information that helps to determine an alternative risk profile with the required feature set. This decision should be the best choice for the organisation's preservation programme. The employment of data mining techniques facilitates this task for a human expert by performing complex calculations and comparisons.

Two scenarios were analysed during evaluation. In the first scenario, we performed the sample risk profile calculation. The hypothesis is that an institutional expert will define some of the most important risk factors and apply them as an input to the data mining tool. The output of the tool should be the given input accomplished with risk factor settings for the remaining risk factors from the nearest expert risk profile. The calculated profile then supports the fuzzy model calculations for endangerment analysis as described in [7].

The second evaluation scenario addresses the visualisation of some risk factor dimensions. The hypothesis for this scenario is that visualisation of particular risk factor dimensions will facilitate and speed up endangerment analysis and demonstrate a level of agreement between important risk factors. Thus, a preservation expert can adjust required risk factor settings in order to reduce preservation risks.

B. Evaluation Data Set

The basis for the risk metrics calculation was provided through an exploratory study organised by Heather Ryan [14] in which digital preservation experts evaluated twenty eight file format endangerment factors (see Table I). Table I represents the dataset as a view from the file format survey data (FFSD) in the first version. In the survey digital preservation experts rated 28 risk factors from 1 to 3, where 3 stands for the high impact of the risk factor and 1 for the low impact. We interpreted the expert ratings of the endangerment factors as levels of risk associated with each factor. The risk estimation ratings from the ten trusted digital preservation experts were evaluated for each of these factors based on their knowledge and expertise. The estimated risk level for each risk factor is depicted in the first column "Risk Level". The columns to the right from the "Risk Factor" column present ten expert risk profiles, whereas the number marks an expert index.

For evaluation purposes well known risk factors were selected and each risk factor was graded in the range from 3 (high impact on preservation risk) to 1 (low impact on preservation risk).

For evaluation of the modified standard score in the visualisation sample two selected expert profiles are used (see Table II). These manually aggregated metrics were used as an input data by the tool for visualisation and analysis of risk factor coherences.

C. Experimental Results and Its Interpretation

The experimental results are presented in two tables. Table II demonstrates the calculated modified standard score values for two selected expert risk profiles with indexes 3 and 5. The associated original values for the expert risk profiles presented in columns "E3" and "E5" can be found in the Table I in the columns "3" and "5".

In the calculated Table II the associated modified standard scores for all twenty eight evaluated risk factors are depicted. The first column is a serial number associated with the X axis in the Fig. 4. The columns "E3" and "E5" contain the original risk factor ratings of the experts 3 and 5. The columns "MSS3" and "MSS5" comprise associated calculated modified standard scores. The last column is a difference between "MSS3" and "MSS5".

Fig. 4 demonstrates visualisation of the values from the Table II. This plot demonstrates the relation between two

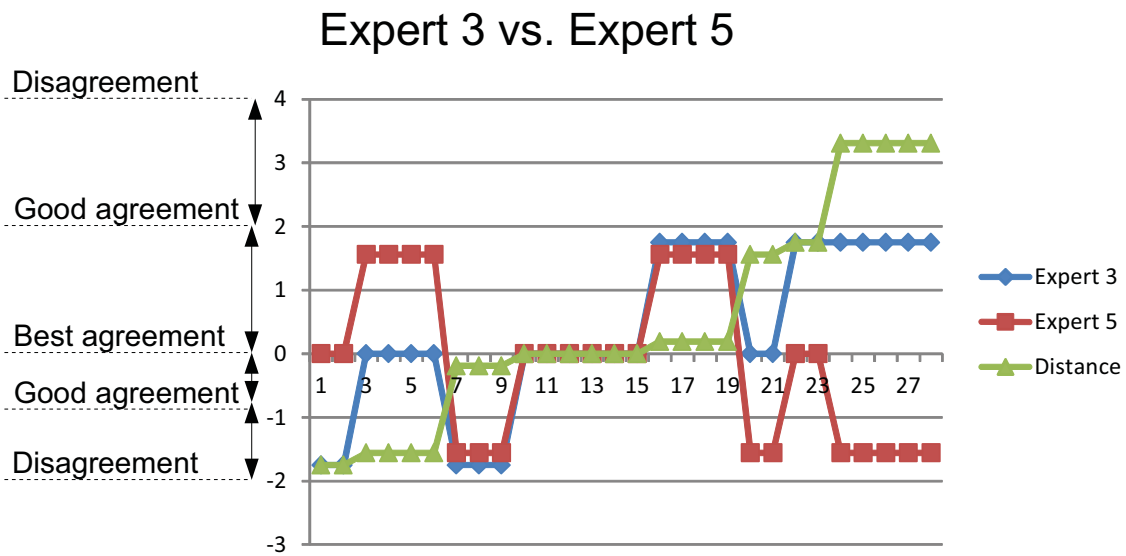

Fig. 4: Plot for relation of risk factor settings between expert 3 and expert 5.

TABLE I: Risk Factors Rating for Digital Preservation of File Formats from the Survey

		ExpertID (EID)									
Risk Level	Risk Factor	1	2	3	4	5	6	7	8	9	10
middle	Value	2	1	2	3	3	1	1	1	3	2
low	Geographical Spread	2	1	3	1	1	1	2	1	2	1
low	Domain Specificity	1	2	3	2	2	2	2	1	2	2
low	Viruses	1	2	1	1	1	2	2	1	1	1
low	Availability Online	2	2	2	3	2	1	1	1	1	1
low	Institutional Policies	2	1	1	2	1	2	2	1	3	2
middle	Specification Quality	2	2	2	3	3	1	3	2	3	2
high	Backward/Forward Compatibility	2	3	3	2	3	3	2	3	2	3
high	Community 3D Support	3	3	2	3	1	2	3	3	3	3
high	Complexity	2	3	2	3	2	2	2	3	3	3
low	Compression	2	1	1	2	2	1	1	2	1	3
middle	Cost	3	3	2	3	2	1	2	3	2	2
low	Developer Support	1	2	3	2	1	1	1	2	2	3
middle	Ease Of Identification	1	2	2	3	3	3	2	2	3	2
middle	Ease Of Validation	1	2	2	3	3	2	2	2	3	3
middle	Error Tolerance	1	2	3	3	3	2	1	3	1	1
high	Expertise Available	2	3	2	3	2	3	3	3	3	3
high	Legal Restrictions	2	3	3	3	1	3	2	2	2	3
low	Life Time	1	3	2	2	1	1	1	2	2	3
low	Metadata Support	1	1	1	2	2	1	1	3	2	2
high	Rendering Software Available	3	3	3	3	3	3	3	3	3	2
low	Revision Rate	1	2	2	2	2	2	2	2	2	3
high	Specification Available	3	3	3	3	3	3	3	3	3	3
middle	Standardization	2	2	3	3	1	2	2	3	2	3
low	Storage Space	1	1	1	1	1	2	1	2	2	1
middle	Technical Dependencies	2	3	3	2	1	3	2	2	3	3
low	Technical Protection Mechanism	3	1	2	2	2	1	1	1	2	3
high	Ubiquity	3	3	3	2	2	3	2	3	2	3

selected expert setting vectors for 28 evaluated risk factors. The risk factor ratings of the experts 3 and 5 are marked by the green and red colour correspondingly. The green graphic demonstrates the difference between expert ratings. The X axis shows 28 risk factors numbered in range from 1 to 28. The associated labels are presented in the Table II. The Y axis is range of the modified standard score. Fig. 4 shows the best agreement experts have in the segment from 7 to 20 on the X axis. In this segment their ratings are either the same (e.g. for risk factors “Availability Online”, “Cost”, “Complexity”, “Expertise Available”) or have a small difference of 0,19 (e.g. for risk factors “Viruses”, “Institutional Policies”, “Rendering Software Available”, “Life Time”). In the segments from 1 to 6 (e.g. “Compression”, “Metadata Support”, “Value”) and 21 to 28 (e.g. “Domain Specificity”, “Ubiquity”, “Standardization”, “Technical Dependencies”) the agreement between experts is not so good. The differences rages from 1,56 to 3,31.

Risk factors for MSS values from expert 3 and 5 correspondingly [1,75;1,56] (“Backward/forward compatibility”, “Error tolerance”, “Rendering software available”, “Specification available”), [0;0] (“Availability online”, “Complexity”, “Cost”, “Expertise available”, “Revision rate”, “Technical protection mechanism”) and

TABLE II: The Modified Standard Scores for Two Selected Expert Risk Profiles.

Nr	Risk Factors	E3	E5	MSS3	MSS5	Distance
1	Compression	1	2	-1,75	0	1,75
2	Metadata Support	1	2	-1,75	0	1,75
3	Value	2	3	0	1,56	1,56
4	Specification Quality	2	3	0	1,56	1,56
5	Ease Of Validation	2	3	0	1,56	1,56
6	Ease Of Identification	2	3	0	1,56	1,56
7	Viruses	1	1	-1,75	-1,56	0,19
8	Institutional Policies	1	1	-1,75	-1,56	0,19
9	Storage Space	1	1	-1,75	-1,56	0,19
10	Availability Online	2	2	0	0	0
11	Cost	2	2	0	0	0
12	Complexity	2	2	0	0	0
13	Expertise Available	2	2	0	0	0
14	Revision Rate	2	2	0	0	0
15	Technical Protection Mechanism	2	2	0	0	0
16	Rendering Software Available	3	3	1,75	1,56	0,19
17	Specification Available	3	3	1,75	1,56	0,19
18	Backward/forward Compatibility	3	3	1,75	1,56	0,19
19	Error Tolerance	3	3	1,75	1,56	0,19
20	Life Time	2	1	0	-1,56	0,19
21	Community 3D Support	2	1	0	-1,56	1,56
22	Domain Specificity	3	2	1,75	0	1,56
23	Ubiquity	3	2	1,75	0	1,75
24	Developer Support	3	1	1,75	-1,56	1,75
25	Legal Restrictions	3	1	1,75	-1,56	3,31
26	Standardization	3	1	1,75	-1,56	3,31
27	Technical Dependencies	3	1	1,75	-1,56	3,31
28	Geographical Spread	3	1	1,75	-1,56	3,31

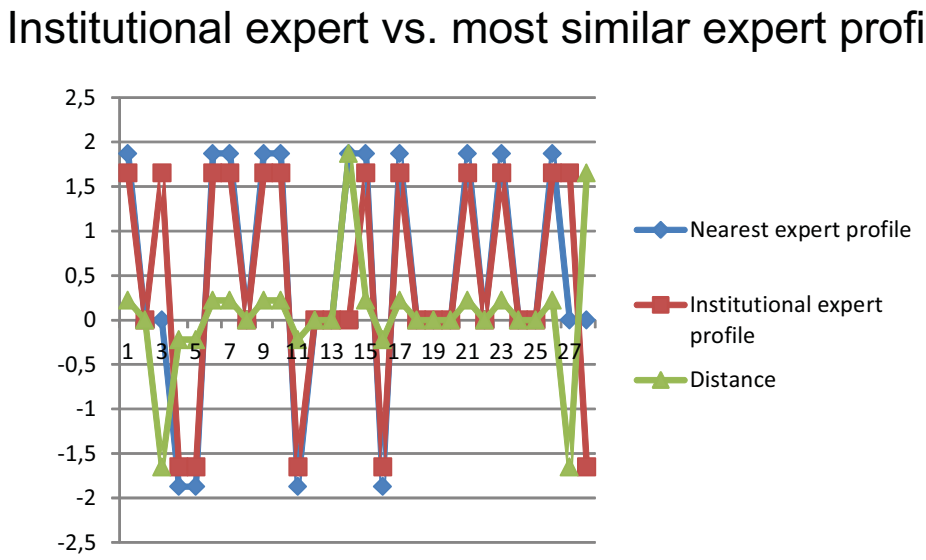

Fig. 5: Plot for relation of risk factor settings between institutional expert and the most nearest expert profile.

[-1,75;-1,56] ("Viruses", "Institutional policies", "Storage space") demonstrate best agreement between experts on these metrics. These modified standard score points correspond with associated initial setting points [3;3], [2;2] and [1;1]. When another risk estimation scale is applyied or different expert employs different scales, these association is not so obvious.

Risk factors "Ubiquity", "Domain specificity", "Compression", "Metadata support", "Value", "Specification quality", "Ease of identification", "Ease of Validation", "Community 3D support" and "Life time" demonstrate good agreement between experts on these metrics.

Risk factors "Geographical spread", "Developer support", "Legal restrictions", "Standardization" and "Technical dependencies" demonstrate disagreement on these metrics.

This approach should support the definition of institutional policies for preservation risk calculation. The knowledge about risks reduces endangerment level of a digital collection. Employing the provided algorithm the institutional expert can either select between predefined expert settings or estimate important risk factors by themself and find the most similar expert profile for the definition of remaining values.

In the first evaluation scenario the institutional expert selects between ten expert profiles. In the second scenario, the institutional expert performs the similarity calculation using the most important definitions as input. Then the resulting risk definition set is one of the ten expert profiles with some overwritten values from the input vector.

For the second scenario the cosine similarity method [11] was applied (see 2). This method returns a cosine similarity between two vectors, where x_i and y_i are vector attributes - risk factors. In the sample case (see Fig. 5) these vectors are risk settings of the institutional expert and the most nearest expert profile (see Table III). We use cosine similarity because test data may be sparse. In the

TABLE III: Results of the Cosine Similarity Calculation.

Risk Factor	Expert	Institution	Recommended	Expert MSS	Inst. MSS
Value	3	-	3	1,87	1,65
Geographical Spread	2	-	2	0	0
Domain Specificity	2	3	3	0	1,65
Viruses	1	-	1	-1,87	-1,65
Availability Online	1	-	1	-1,87	-1,65
Institutional Policies	3	3	3	1,87	1,65
Specification Quality	3	-	3	1,87	1,65
Backward/forward Compatibility	2	-	2	0	0
Community 3D Support	3	-	3	1,87	1,65
Complexity	3	-	3	1,87	1,65
Compression	1	-	1	-1,87	-1,65
Cost	2	-	2	0	0
Developer Support	2	-	2	0	0
Ease Of Identification	3	2	2	1,87	0
Ease Of Validation	3	-	3	1,87	1,65
Error Tolerance	1	-	1	-1,87	-1,65
Expertise Available	3	-	3	1,87	1,65
Legal Restrictions	2	-	2	0	0
Life Time	2	-	2	0	0
Metadata Support	2	-	2	0	0
Rendering Software Available	3	-	3	1,87	1,65
Revision Rate	2	-	2	0	0
Specification Available	3	-	3	1,87	1,65
Standardization	2	-	2	0	0
Storage Space	2	-	2	0	0
Technical Dependencies	3	3	3	1,87	1,65
Technical Protection Mechanism	2	3	3	0	1,65
Ubiquity	2	1	1	0	-1,65

TABLE IV: The Groups of the Expert Risk Profiles.

Risk Profile Group	Expert ID	SUM of expert raitings
Low	1	52
Low	7	52
Low	5	54
Low	6	54
Middle	2	60
Middle	8	60
Middle	3	62
Middle	9	63
High	10	66
High	4	67

given example, the institutional expert believes that important high level risk factors are “Domain specificity”, “Institutional policies”, “Technical dependencies” and “Technical protection mechanism”. The middle level risk factor would be “Ease of identification”. Finally, the low level risk factor for this particular institution is “Ubiquity”. Remaining values are not defined in the institutional policy. Applying the cosine similarity search to the expert knowledge pool calculates that the most similar expert vector has index nine. Therefore, the resulting risk profile for the given institution is evaluated by merging the institutional profile with the found expert profile and is presented in the Table III. The column “Expert” stands for the expert profile. The column “Institution” means institutional profile settings. The next column “Calculated” means intitutional profile merged together from the original institutional risk profile and the nearest expert risk profile. The column “Expert MSS” contains calculated modified standard scores for the expert risk profile. And the column “Inst. MSS” comprises associated modified standard score values for the merged institutional risk profile. The initial institutional risk profile is presented as a multidimensional vector [0, 0, 3, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 1], where 3 stands for high risk level, 2 for middle risk level, 1 for low risk level and 0 means “not set”. Calculating of the cosine similarity between the institutional risk profile and ten expert risk profiles produces ten associated cosine similarity values [0.4279539686731926, 0.3619009099557445, 0.43069040314741897, 0.3799574797940304, 0.3676220091661415, 0.4665971654801027, 0.40204576633284433, 0.2843507149652278, 0.47667878002211034, 0.44060109511232537]. The maximal cosine similarity 0.476678780022 and therefore, the best match to the institutional profile has expert number nine. His risk profile vector is [3, 2, 2, 1, 1, 3, 3, 2, 3, 3, 1, 2, 2, 3, 3, 1, 3, 2, 2, 2, 3, 2, 3, 2, 2, 3, 2, 2]. Merging these two risk profiles we produce the resulting institutional risk profile depicted in Fig. 5. The most presented risk factors [1,87;1,65], [0;0] and [-1,87;-1,65] have best agreement. Some of the risk factors [0;1,65], [0;-1,65] and [1,87;0] (Domain specificity, Ease of identification and Ubiquity) are slightly different but still have good agreement. The green “Difference” line in Fig. 5 is mostly near 0 that graphically demonstrates best agreement between profiles.

The Table IV demonstrates the possible risk profile groups for the given expert group (see the first column “RP Type”). Having the autocompleted risk profile from the previous evaluation step we use it as an input in order to estimate matching risk profile group by employing of the naive Bayes method. The next columns represent ratings for 28 risk factors from the survey by their number, which are correlated with the labels from the Table III. The column “EID” comprises the expert IDs for expert identification. And the “SUM” column contains overall count of ratings for the associated expert.

For risk profile group evaluation we calculate probabilities:

- $p(L|rf_i \in R)$
- $p(M|rf_i \in R)$
- $p(H|rf_i \in R)$.

Having these three probabilities we take the highest probability as a result.

The resulting computation by means of naive Bayes algorithm for the given input returns the risk profile group “Middle”. That means that the given at the beginning of evaluation institutional risk factor values most likely belong to the moderate risk profile group. Having this information an institutional expert can adjust risk factors in order to change the risk profile group. E.g. if an expert expects more endangerment risks he would like to have the “High” risk profile group.

These results demonstrate (see Fig. 4 and 5) that a semi-automatic approach for risk factors visualisation is very effective and it is a signicant improvement compared to manual analysis.

V. Conclusion

In this work we presented an approach for the easy creation and classification of an institutional risk profile for endangerment analysis of file formats.

The main contribution of this work is the employment of data mining techniques to support risk factors set up with just a few of the most important values for a particular organisation. The resulting risk profile and its group is used to support digital preservation experts with semi-automatic estimation of endangerment level for file formats.

The presented method employs a domain expert knowledge base aggregated from a survey in order to detect preservation risks for particular institution.

Another contribution is support for the visualisation and analysis of risk factors for requried dimension. To facilitate easier evaluation, the aggregated information about the risk factors is presented as a multidimensional vector. The proposed methods improve the visibility of risk factor information and the quality of a digital preservation process.

We make use of data mining techniques like the modified standard score method in order to analyse aggregated data and the cosine similarity calculation in order to compare risk profiles. The employment of the naive Bayes algorithmus classifies the selected institutional risk profile and makes recommendation to an expert regarding the most likely risk profile group.

In the evaluation section, different risk factor dimensions are exposed. The presented plot demonstrates coherences in risk factors and help in solving practical digital preservation issues. Using the developed approach and adjusting input data, experts have the ability to choose the appropriate risk factor setting for digital preservation planning in their institution.

The presented approach is designed to facilitate decision making with regard to preservation of digital content in libraries and archives using domain expert knowledge. As future work we plan to increase the amount and quality of aggregated expert information and to extend the tool with additional visualisation scenarios.

References

[1] P. Ayris, R. Davies, R. McLeod, R. Miao, H. Shenton, and P. Wheatley. The life2 final project report. Final project report, LIFE Project, London, UK, 2008.

[2] L. C. David Tarrant, Steve Hitchcock. Where the semantic web and web 2.0 meet format risk management: P2 registry. *International Journal of Digital Curation*, 6(1):165–182, 2011.

[3] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny. Cosine similarity scoring without score normalization techniques. *in Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop (Odyssey 2010)*, pages 71–75, 2010.

[4] S. Gordea, A. Lindley, and R. Graf. Computing recommendations for long term data accessibility basing on open knowledge and linked data. *Joint proceedings of the RecSys 2011 Workshops Decisions@RecSys'11 and UCERSTI 2*, 811:51–58, November 2011.

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

[6] R. Graf and S. Gordea. A risk analysis of file formats for preservation planning. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres2013)*, pages 177–186, Lissabon, Portugal, Sep 2013. Biblioteca Nacional de Portugal, Lisboa.

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

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

[9] J. Hunter and S. Choudhury. Panic: an integrated approach to the preservation of composite digital objects using semantic web services. *International Journal on Digital Libraries*, 6, (2):174–183, September 2006.

[10] A. N. Jackson. Formats over time: Exploring uk web history. *Proceedings of the 9th International Conference on Preservation of Digital Objects*, pages 155–158, October 2012.

[11] A. Karnik, S. Goswami, and R. Guha. Detecting obfuscated viruses using cosine similarity analysis. In *Modelling Simulation, 2007. AMS '07. First Asia International Conference on*, pages 165–170, March 2007.

[12] G. W. Lawrence, W. R. Kehoe, O. Y. Rieger, W. H. Walters, and A. R. Kenney. Risk management of digital information: A file format investigation. june 2000.

[13] D. Pearson and C. Webb. Defining file format obsolescence: A risky journey. *The International Journal of Digital Curation*, Vol 3, No 1:89–106, July 2008.

[14] H. Ryan. File format study. *School of Information and Library Science, University of North Carolina at Chapel Hill*, 2, 2013.

[15] D. Tanner. Using statistics to make educational decisions. *Library of Congress Cataloging-in-Publication Data*, pages 77–104, 2012.

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

[17] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[18] J. Ye. Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Mathematical and Computer Modelling*, 53(1?2):91 – 97, 2011.

[19] R. Zacharski. *A Programmer's Guide to Data Mining: The Ancient Art of the Numerati*. 2012.

[20] H. Zhang. The Optimality of Naive Bayes. In V. Barr and Z. Markov, editors, *FLAIRS Conference*. AAAI Press, 2004.