

A Model for Format Endangerment Analysis using Fuzzy Logic

Roman Graf
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
roman.graf@ait.ac.at

Sergiu Gordea
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
sergiu.gordea@ait.ac.at

ABSTRACT

This paper presents an approach for merging information automatically aggregated from open repositories and expert knowledge related to digital preservation. The main contribution of this work is the employment of fuzzy models to support digital preservation experts with semi-automatic estimation of “endangerment level” for file formats. Our goal is to make use of a solid knowledge base automatically aggregated from linked open data repositories to detect conflicts and inaccuracies in this data in order to improve the quality of a risk analysis process. The proposed method is meant to facilitate decision making with regard to preservation of digital content in libraries and archives using knowledge of domain experts. To allow reasoning even in case of inconsistent data we employ fuzzy logic techniques for transforming information about formats in user friendly metrics. The goal is to bring conflicting and incorrect information to the surface for correction and improvement by community. The analysis of a survey regarding the risk factors for file formats was used as an input for fuzzy model and is presented in evaluation part.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: System issues; H.3.5 [Online Information Services]: Web-based services

Keywords

digital preservation, risk analysis, linked open data, preservation planning, ontology matching, information integration

1. INTRODUCTION

In the past years, libraries, archives and museums have been carrying out large-scale digitization projects. As result, new digital collections that comprise millions of objects were created; and the goal is to make them available on long term basis. Consequently, digital libraries are facing a paradigm shift regarding preservation, maintenance and quality assurance of these collections. Therefore, automated solutions for data management and digital preservation are imperatively necessary.

The core of preservation planning activities deals with the evaluation of appropriate formats used for encoding digital content. The preservation risks for a particular file format are difficult to estimate [5]. The definition of risk factors and associated metrics is still an open research topic in digital preservation community¹. Involvement of digital preserva-

tion experts is required for collecting complete information and evaluating preservation risks[1]. Currently, each institution defines its own risk factors for long term preservation depending on particular project, preservation goals, workflows and assets. The richness and the quality of individual knowledge bases play an important role in taking decisions on preservation planning, but often these resources do not contain all necessary semantic information for performing a faithful (automatic) evaluation of file formats.

Many file formats are properly documented, are open-source and well supported by Software Vendors. Other formats may be outdated or no longer functional with modern software or hardware. There are also custom/proprietary formats, which might be obsolete and not renderable with commodity hardware. To address these problems, we employ the File Format Metadata Aggregator (FFMA) [4] system and the information integration approach depicted in Figure 1. FFMA is a part of knowledge base recommender DiPRec [3], which reuses the experience of building preservation planning tools and offers assessment for long-term preservation of digital content. This tool performs an analysis of file formats based on the concept of risk scores.

The main contribution of the current work is the development of an Expert System based on fuzzy rules for performing the analysis of digital collections. Fuzzy rules are employed for handling the level of uncertainty associated to the information aggregated from LOD. Decision support based on the elaborated rule engine provided by FFMA and fuzzy rules is meant to support institutions like libraries and archives with assessment for analyzing their digital assets. The valuable basis for risk metrics calculation was provided by survey organised by the University of North Carolina[6] which takes in account twenty eight risk factors. Evaluation metrics were defined for each of these factors based on the knowledge of digital preservation community. We aim at defining a fuzzy model and metrics intended to provide decision making support based on expert community knowledge. The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains risk analysis process, knowledge aggregation process from LOD repositories and covers also ontology mapping, fuzzy modelling and algorithmic details of endangerment analysis. Section 4 presents the experimental setup, file formats survey, applied methods for fuzzy analysis and results. Section 5 concludes the paper and gives an outlook about planned future work.

¹<http://www.openplanetsfoundation.org/blogs/2013-09-30-assessing-file-format-risks-searching-bigfoot>

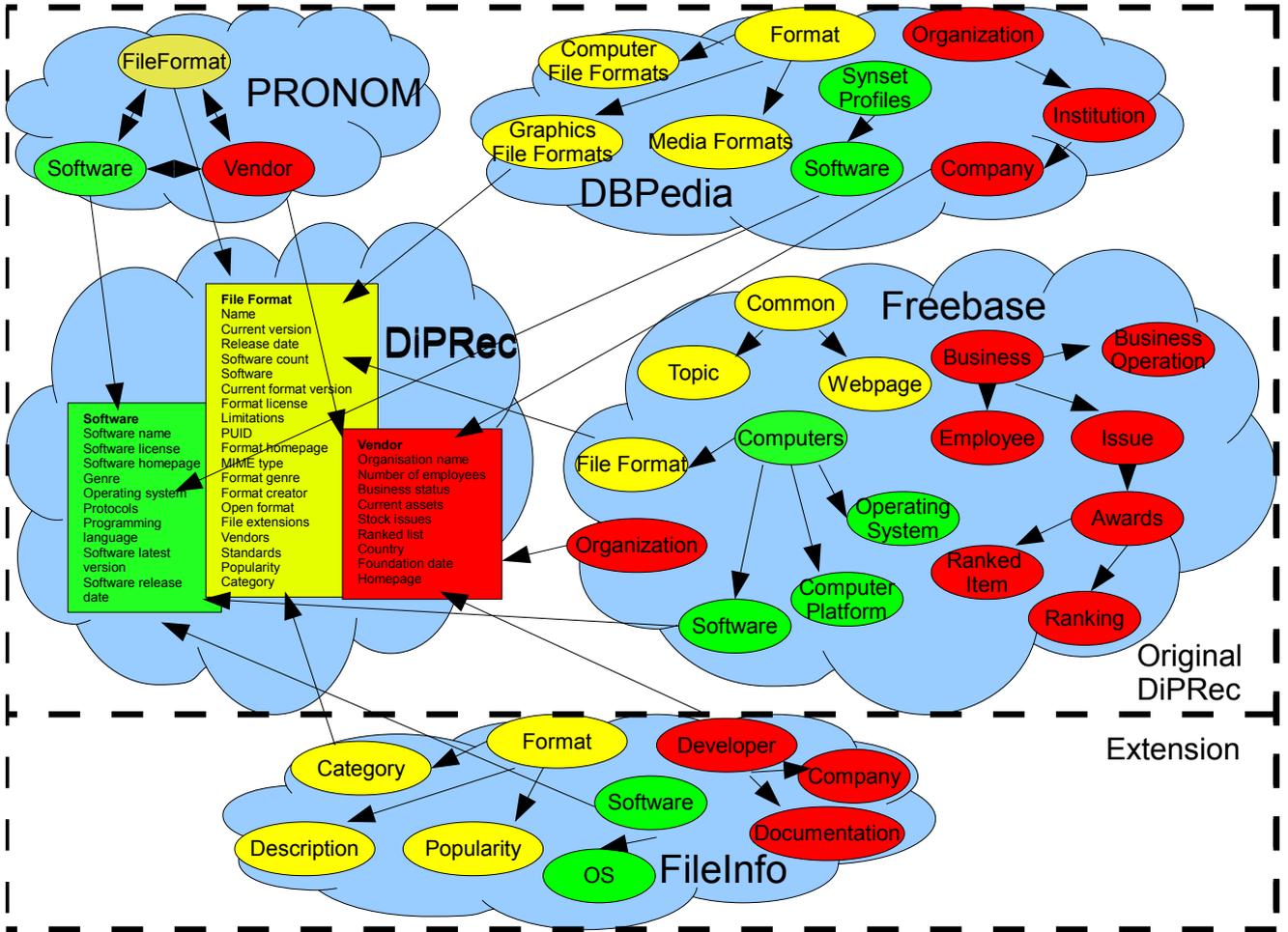


Figure 1: PRONOM, DBPedia, Freebase and Fileinfo digital preservation domain related ontology sections mapped to the DiPRec file format ontology.

2. RELATED WORK

The main issue addressed in this work is the controversial understanding of format obsolescence. Good overview on this topic gives Andrew Jackson's research in [8] where he has evaluated competing hypotheses regarding software obsolescence issue. He employed format identification tools for selecting appropriate preservation strategies. One of these hypothesis is presented by Rothenberg [12] and emphasizes that all formats should be considered brittle and transient, and that frequent preservation actions will be required in order to keep data publicly accessible. In contrast to that hypothesis the Rosenthal [11] claims that no one supporter of format migration strategy was able to identify even one format that has gone obsolete in the last two decades. Rosenthal argues that the network effects of data sharing inhibit obsolescence.

An accurate format identification and rendering is a challenging task due to malformed MIME types, rendering expenses, dependence on some content not embedded in the file, missing colour table, changed fonts, etc. In [8], the

author examines how the network effects could stabilise formats against obsolescence in order to understand the warnings, choices and costs involved. This evaluation should help to meet preservation strategy: either to perform frequent preservation actions to keep data accessible or to concentrate on storing the content and using available rendering software. The result of evaluation demonstrates that most formats last much longer than five years, that network effects stabilise formats, and that new formats appear at a modest, manageable rate. However, he also found a number of formats and versions that are fading from use and that every corpus contains its own biases.

The digital preservation tools like PANIC [7], AONS II [10], SPOT [14], P2 registry [2], aimed at identifying file formats used for encoding digital collections and informing repository managers of events that might impact the access to the stored content. They also define mechanisms for alerting when file formats become obsolete. These tools demonstrate significant differences to our approach. They do not apply metrics for risk calculation, and take in account significantly less properties. Often these properties are estimated

and not measurable, do not exploit the knowledge available to the public, or are limited to particular open sources. Also there is no common understanding in the community about the meaning of term “obsolete” as mentioned above. In the proposed approach we do not intend to mark down obsolete formats, since there are different hypotheses and no common accepted definition for format obsolescence. We estimate obsolescence in relation to the additional effort required to render a file beyond the capability of a regular PC setup in particular institution. This is consistent with the “institutional obsolescence” concept saying that a particular format that would not render anymore on a PC in an institution’s reading room should be considered as obsolete.

An application of natural language words instead of numbers for computing and reasoning using fuzzy logic is described in [9]. A survey of the fuzzy logic controller (FLC) presented in [15] evaluates a linguistic control methodologies, the derivation of the fuzzy control rules and an analysis of fuzzy reasoning mechanisms. The qualitative safety modelling in [13] is performed employing fuzzy IF - THEN rules. Compared to existing digital preservation recommenders the proposed approach is more effective due to the use of more complex fuzzy rules. Existing tools are not well suited for dealing with aggregated LOD data having a level of uncertainty due to conflicts and inaccuracies between different sources. Inaccuracies in this sense are slightly different measurements, which do not impact the overall evaluation of the risk factor. E.g. software count for PDF format provided by Freebase is 12 whereas Fileinfo describes 25 tools. We define conflicts as significant contradictions implying different conclusions on risk factor evaluation. E.g. PRONOM classification for PDF format is “page description” that contradicts to the Freebase genre for this format “graphics file format”. A fuzzy-logic-based approach is more appropriate for the correctness analysis. The provided Expert System deals directly with the linguistic terms commonly used in digital preservation community for quality assessment. Our research focuses on the development and representation of user friendly and easy understandable linguistic variables to confidence levels. These variables are then quantified using fuzzy logic. Inspired by [10] we realized the need to develop a central web service that shares the results of open data aggregation and correctness assessments with the community of interest. We aim at defining endangerment metrics based on experience of community members which share their individual expertise on defining and identifying risk factors.

3. ENDANGERMENT ANALYSIS

Digital preservation is an area where we have to take into account fuzziness and a high amount of descriptions regarding the encoding formats. The description of file formats aggregated from open repositories is often far from being complete and accurate. Therefore, we support the aggregation of expert knowledge for enhancing such a repository with high confidence information. The proposed Expert System should identify conflicts and inaccuracies and provide assessment on the “institutional obsolescence” of file formats. We realized that the digital preservation community already uses the multiple format registries and doesn’t trust “expert systems” when making preservation related decisions. Instead, they recognize the need for support systems that aggregate and compare knowledge about the file formats (i.e.

in form of metrics). This approach should help to uncover conflicting and not trusted information so that domain experts may correct it according to the policies established in their institution.

3.1 Endangerment Computation Model

The final conclusion of the rule-based system estimates the endangerment level (i.e. high vs. middle vs. low) for the analysed file formats. The computation of the overall endangerment level is performed by integrating the view of expert community (see Figure 2) and by using the associated fuzzy rule model (see Figure 3). The Endangerment Computation Model (ECM) can be customized to model the policies of a particular organisation.

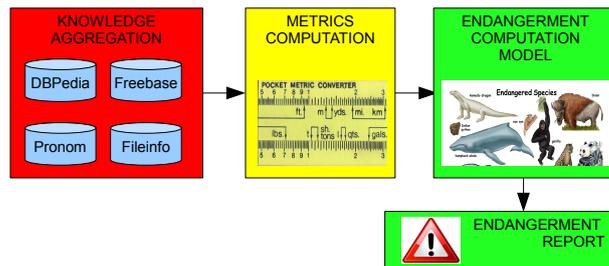


Figure 2: The workflow for the format endangerment and correctness analysis.

Figure 2 sketches the workflow used within the endangerment analysis process. The building of the knowledge base (i.e. Knowledge Aggregation) is a prerequisite for performing the endangerment computations. This includes the acquisition of expert knowledge and the aggregation of file format data in a common domain model. The creation of endangerment analysis reports is a two-step process based on the definition of fuzzy factors (i.e. Endangerment Computation Model) and the computation and interpretation of fuzzy metrics (i.e. Metrics Computation). The final endangerment report contains detailed information about the endangerment level, including of quantifications of the evaluation factors, the computed metrics for inaccuracy and conflicting descriptions of each format.

The model proposed for evaluating the endangerment level comprises three rule blocks grouped by their impact level (see Figure 3). Each of the factors taken in account are evaluated basing on the associated metrics. The analysis of risk factors calculations delivers three folded results. An “endangerment” output estimates the endangerment levels. A “conflicts” output analyses the conflicting information received from different sources. This analysis takes in account format properties that include description, software count, vendor count, compression, versions count, existence period, complexity, dissemination, deprecation, genre, homepage, standard, migration, digital rights, popularity, web browser support, MIME, timestamp, etc. This module estimates the severity of the conflicts and their occurring rate. Finally, we have defined the “inaccuracies” part that notices inaccuracies associated with particular file format, it estimates their severity level and their count. By combining the outputs of these three modules, the inference engine concludes about

the overall endangerment level and evaluates the risks for the analysed format.

The goal of the fuzzyfication is to map the numerical values to the decision variables by using the membership functions. By combining all defined fuzzyfied variables we can construct a hierarchical fuzzy inference system, since the output of a fuzzy inference module can be used as input for the next level of inference within the system. As example, the inference module for the complexity risk factor depicted in Figure 4 is used as input for the inference model presented in Figure 3. A concrete example of complexity calculation is presented in Section 4. A fuzzy set estimates the “Endangerment Level” of a factor as belonging to the impact categories “Low”, “Middle” and “High”. This is decided by using membership functions as the ones presented within the Equations 1-5.

$$(U, m) = \left\{ \frac{m(x_{LOW})}{x_{LOW}}, \frac{m(x_{MID})}{x_{MID}}, \frac{m(x_{HIGH})}{x_{HIGH}} \right\}, \quad (1)$$

$$x \in U \quad (2)$$

$$m(x_{LOW}) = \begin{cases} 1, & \text{if } 0 < x \leq 25, \\ -\frac{x}{10} + 3,5, & \text{if } 25 < x \leq 35, \end{cases} \quad (3)$$

$$m(x_{MID}) = \begin{cases} \frac{x}{10} - 2,5, & \text{if } 25 < x \leq 35, \\ 1, & \text{if } 35 < x \leq 55, \\ -\frac{x}{10} + 6,5, & \text{if } 55 < x \leq 65, \end{cases} \quad (4)$$

$$m(x_{HIGH}) = \begin{cases} \frac{x}{10} - 6,5, & \text{if } 55 < x \leq 65, \\ 1, & \text{if } 65 < x \leq 100. \end{cases} \quad (5)$$

Where (U, m) denotes a fuzzy set U with membership function $m(x)$. The concrete instances x belong to the set U with different degrees of membership quantified in numeric values - from not included ($m(x) = 0$) to fully included ($m(x) = 1$). The Figure 6 depicts graphical representation of previously defined fuzzy rules and their membership functions.

3.2 Metric Computation Model

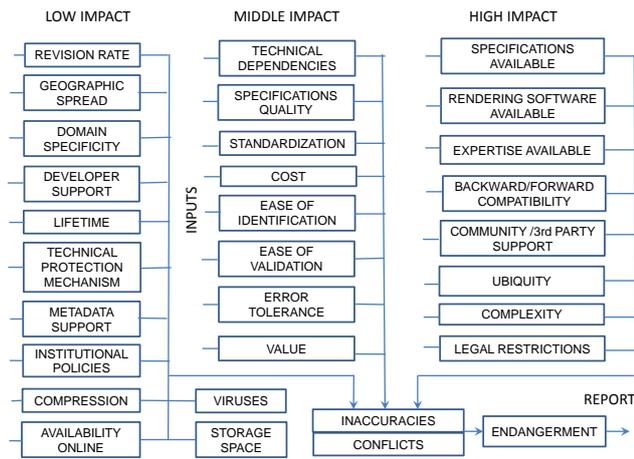


Figure 3: An inference model for calculation of endangerment level.

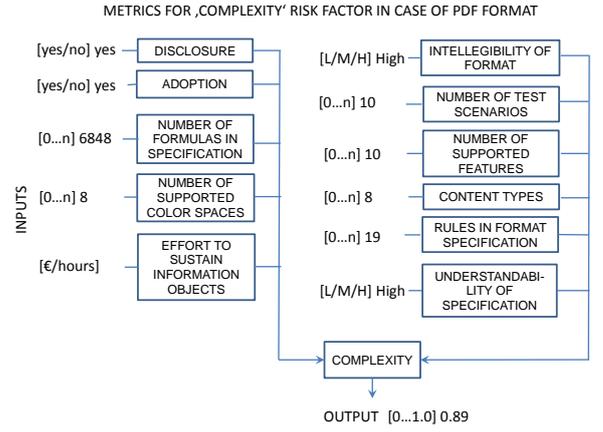


Figure 4: An inference system for calculation of the complexity risk factor by employing of the associated metrics for the given file format.

The metrics for the rule “Complexity” in Figure 4 have different ranges for input values that are presented in angular braces. These ranges can be numerical, boolean or textual. The input values for these ranges can be retrieved from LOD repositories employing FFMA tool. As a sample for this rule we will analyze the PDF format. The metric “DISCLOSURE” becomes input value “yes” since it is an open standard ISO 32000 as stated in “Adobe” vendor documentation pointed by Fileinfo registry. This format is broadly used by thousands of vendors worldwide. The estimation of document numbers is hard to define because of different types of documentation like books, textual documents and HTML tutorials. We have counted 1662 tutorial documents and each of them has in average 2 pages. Number of formulas in documentation has low relevance in our opinion but it would make sense to estimate number of code snippets or screenshots. In that sense we counted this metric with 4 per page in average. Features count can be also found in documentation and is given by at least 10 top features but that can’t be automated. We have found 8 color spaces. The effort to sustain information objects can be very different depending on organisation goals and can be measured in money amount and/or working hours. The intelligibility and understandability of this format is high since it can incorporate another formats, renders on different operation systems and has a lot of communities and vendor support. PDF is supported by 28 software tools (see Table 2) that has middle level in our classification. As a part of training we found 10 test scenarios. PDF supports text, drawings, videos, audio, 3D maps, full-color graphics, photos and business logic. Rules of the format is very difficult to estimate since definition of rule is vague. We found 19 rules meaning different aspects of the standard.

The Figures 5 shows an example fuzzy rule with associated values. These example demonstrates membership function $m(x)$ definition.

3.3 Knowledge Aggregation

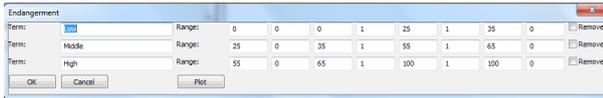


Figure 5: Example fuzzy rule definition for endangerment rule.

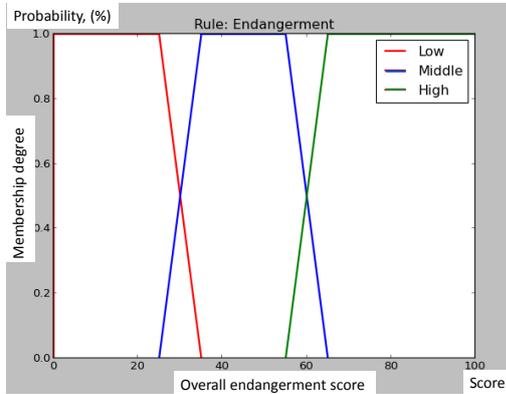


Figure 6: Plot of resulting endangerment level estimation as a result of all factors calculated by associated metrics.

The FFMA module for aggregation of file format descriptions collects information from LOD repositories and enhances it by aggregation of expert knowledge. A specific exploitation context may customize which LOD repositories should be used and which file format properties are of interest for particular institutional context. The File Format Data Aggregation module is responsible for collecting descriptions on file format-related information from the open knowledge bases, while the FFMA engine combines the outcome of the module with the knowledge manually provided by domain experts. The acquired domain knowledge is stored in a local database and further used for reasoning in risk computation process. The external knowledge sources like DBPedia and Freebase manage huge amounts of LOD triples, which allows one to extract fragmental descriptions on file formats, software applications and software vendors.

4. EXPERIMENTAL EVALUATION

The goal of evaluation of format risks was the enhancement of FFMA knowledge base and validation of aggregated data. This process is described in the correctness calculation workflow (see Figure 2). Our hypothesis is that file format data automatically aggregated from LOD repositories will provide the fuzzy inference engine with valuable information and will enable correctness estimation for different file formats. The “high” confident marked formats should indicate the currently most reliable file formats for digital preservation workflows. A Web service was developed that automatically retrieves file format related data from LOD repositories and performs reasoning on collected information employing specified risk factors. The collected information is processed, normalized, integrated into the knowledge base.

The programming interface of this service supports querying for descriptions of the file formats, software, vendors and associated information. Service supports checking of availability of the information in the service database and retrieving data from LOD repositories if necessary. Another goal of our evaluation is the need to recognise that format is becoming obsolete and prepare adequate preservation planning, strategies and actions in response. Our approach should give an organisation a basis at hand that helps to choose a particular format and renderer. This basis should give a hope that the decision has been made on the basis that that is the best choice for the organisation and for the organisation’s preservation programme. The improvement of Fuzzy technique in comparison to FFMA[5] approach is that fuzzy approach is more flexible and emulates a human expert by concept of partial truth, whereas FFMA risk system knows only True/False modes of truth.

4.1 Evaluation Data Set

For evaluation purposes a subset of 13 representative, well known file formats was selected. The *GIF*, *PNG*, *JPG*, *BMP* and *TIF* formats belong to the raster graphics genre. *MP3* is the most used audio format, while the *PDF* format is mostly used for document formats, having multiple versions and being well supported by Adobe Acrobat toolset. The *HTML* format also has multiple versions and is used for creation of Web pages. The *DOC* and *PPT* are Microsoft formats supporting creation of multimedia documents and presentations. Some outdated file formats are presented by *MAC*, *SXW* and *DXF*. The *MAC* is a bitmap graphic format for the Macintosh, one of the first painting programs for this OS, supporting greyscale-only graphics. The *SXW* is an outdated text format for OpenOffice, while *DXF* is a vector graphic format for AutoCAD.

4.2 Computation of Risk Factors

The previously defined rules should be organized in order to process input values and to infer appropriate conclusions. As an example, the rule-based system may start endangerment identification for PDF format with the inference engine of the “Complexity” factor in Figure 4 which comprises 11 fuzzy preconditions. The particular input values are depicted by the rectangles sorted by impact level that was evaluated from the survey. Having input values on the left side and running calculations we receive a confidence level value 0.89 on the output. According to our FLC definitions depicted in Figure 6 that means that resulting confidence level is “high”. The value “high” is a result of matching the numerical output value 0.89 to the fuzzy rule for calculation of confidence level using member functions in Equation 1, where “low” is defined for values in range from 0 to 0.15, “middle” from 0.14 to 0.88 and “high” from 0.67 to 1.0 respectively. Therefore, the input value of the “Complexity” factor in Figure 3 is 0.89. The Expert System suppose complexity level of the format is “high” if most of the metrics after fuzzification produce total output value greater then 0.67. Each of the metrics can again be formulated as a fuzzy rule according to preferences of particular institution. Fuzzifying this value we map it to the associated numerical value using FLC input variables definition. Aggregating all rule outputs we defuzzify output value of the total endangerment level that is “high” and map it to the resulting number 0.93.

An input variable “Resulting Risk” contains three membership functions flagged by the linguistic variables “Low, Middle and High”. A corresponding graphical representation is shown in Figure 6. The values for these linguistic variables range from 0 to 1 and are coming from the inference engine. For simplicity we transform these values to percents. Therefore, format risk can be defined as high if its value matches in a range between 67 and 100 percent. In contrast middle risk values are between 14 and 88 percent. Finally values between 0 and 20 percent indicate that there is low risk for analyzed file format.

Table 1 demonstrates the results of risk factor rating for file formats provided by a file format study of Ryan Heather. The survey was performed among 21 digital preservation experts in three rounds. The relevance of particular factor from the left column on file format risk is defined by values from 1 to 3. Value 3 in this table stands for “Very relevant”, 2 for “Somewhat relevant” and 1 for “Not relevant at all” respectively. The most relevant factors according to evaluation are listed first. The column “SUM” depicts the sum of all votes. The average relevance per factor was calculated and depicted in the “AVG” column. Also the total endangerment value for each factor was calculated and presented in the column “Endangerment level”. This row demonstrates how relevant is mentioned factor for the whole format estimation by associated linguistic values in range between “Middle” and “High”. That means that the international expert community in digital preservation domain pays significant attention to these factors.

The suggested factors cover most of the risk factors identified in FFMA. Merging these two sets we get a basis for fuzzy system. The main conclusion from the review presented in Tables 1 and 2 is that there is a need for some metrics describing file formats. Such metrics can be automatically provided by extended FFMA risk model[5]. By metrics definition we will stick by previously presented in FFMA and in survey simple range Low/Middle/High. The goal by defining metrics is to automate an evaluation of file format risk. In some situations many metrics probably are not realistic since no universal standards for them exist but nevertheless automation can be possible for institutional use cases with good documented workflows. Estimation of risk factor risks is impossible without definition of quality metrics and relevant semantics.

The ‘Backward/Forward Compatibility’ factor makes influence on how easy and cheap content in original format can be accessed, migrated and meaningfully rendered and is a mitigating factor of endangerment or obsolescence of a file format. Measuring of this factor employs information about software that fails in reading an older format, about font substitution failures and about automatically adjusting the color space. Another attributes for this factor are well documented format specification, rendering software number and documentation, licence management, number of versions, release notes and direct testing support measurement of backward compatibility that should be verified by a human. The ‘Community/3rd Party Support’ factor enables for people to implement open format if it is popular enough. The existence of multiple independent implementations using the same format insures that the format is stable and

well-defined. It can be measured by number of communities, by number of software applications supporting it, by trends of software support compared to previous time period, by emulation environments and by counting the number of users or files. Proprietary formats can not be supported by community. This factor depends on how much of the specifications are published and if a file format contains patented parts or techniques. The ‘Complexity’ factor can have different meaning for different institutions. The level of complexity for e.g. PDF raises so much that the costs of providing access might become unsustainable. Measurement of complexity requires accurate generation of representation network, which is difficult to automate. It is dependent on specifications quality, implementations number for the same functionality within a document, number of testing scenarios. Optionally supported features complicate evaluation of compatibility. The feature rich specification such as JPEG2000 is more complex than a very much reduced specification such as a GIF file. In a long term preservation strategy it can be much harder to migrate a highly complex file format. Complexity attributes are depicted in Figure 4. The ‘Compression’ factor depends on the compression algorithm. If it is public and its implementation well documented and freely available it is better than proprietary compression. It reduces storage needs. Lossy compression could compromise the fidelity or essence of the electronic record. The support for compression does not necessarily cause a file format to become endangered if it is good documented. It is dependent on public availability, application characteristics, lossy/lossless feature. The ‘Cost’ calculation factor for long-term storage is often underestimated and can be measured by the number of hardware and software dependencies for file rendering now and in the future. If there is a high cost associated with processing or accessing files in particular formats, they are less likely to be successfully preserved over time. This factor is linked to the value of the contents of the file. There are costs to implement a production system for making files, for a system to manage them e.g. migration, to train people to work with content, and to transmit via network or to store it. It is dependent on the online access by web browser, licence fee and emulation software. The factor ‘Developer/Corporate Support’ could describe whether the agency which originally created the software/file format is still supporting it, have license or lets some open source agency handle maintenance according to open standard. This factor can be measured by supporting vendors, standardization, implementations number, community governance, number of licences, adoption and documentation quality. The factor ‘Ease of Identification’ is related to automated identification and helps to assess the information object and what is required to sustain it. There are a number of formats today that automated file format identification tools fail to identify. Attributes for this factor are headers, characterisation software such as FIDO, JHOVE, DROID, registries, descriptions and documentation. The factor ‘Ease of Validation’ demonstrates that rendering software (e.g. for MPEG4) is prone to validation errors if it is not conforming to specification. Validation is only going to be available for the most widely used formats, because validation engines are expensive to create. Valid HTML and CSS renders very differently on different platforms. It is important to make the difference between broken files and incompatibility with the rendering software. This factor is

Table 1: Risk factors rating for digital preservation of file formats from the survey

Risk Factor	SUM	AVG	Experts Number	Endangerment Level
Specifications Available	33	3.000	11	high
Rendering Software Available	32	2.909	11	high
Expertise Available	30	2.727	11	high
Backward/Forward Compatibility	29	2.636	11	high
Community/3rd Party Support	29	2.636	11	high
Ubiquity	29	2.636	11	high
Complexity	27	2.455	11	high
Legal Restrictions	27	2.455	11	high
Technical Dependencies	26	2.364	11	middle
Specification Quality	23	2.300	10	middle
Standardization	25	2.273	11	middle
Cost	25	2.273	11	middle
Ease of Identification	24	2.182	11	middle
Ease of Validation	24	2.182	11	middle
Error-tolerance	22	2.091	11	middle
Value	20	2.000	10	middle
Revision Rate	21	1.909	11	low
Geographic Spread	19	1.900	10	low
Domain Specificity	19	1.900	10	low
Developer/Corporate Support	20	1.818	11	low
Lifetime	20	1.818	11	low
Technical Protection Mechanism	20	1.818	11	low
Metadata Support	18	1.636	11	low
Institutional Policies	16	1.600	10	low
Compression	17	1.545	11	low
Availability Online	15	1.500	10	low
Storage Space	15	1.364	11	low
Viruses	13	1.300	10	low

dependent on checksum, file headers and degree of non conformance. The factor 'Error-tolerance' is a money issue since even one bit can break rendering. Preservation repositories commonly use bit preservation techniques (e.g. replication) to be able to ascertain when and where bit corruption occurred. By information loss, lost information can be often recovered if the loss is small. Error-tolerance can be measured by testing, popularity, information loss and error rate. The factor 'Expertise Available' impacts the long-term viability of the rendering, migration or emulation. A digital preservation expert needs to understand the whole platform especially by proprietary formats. The attributes for expertise estimation are expert skill level, experience, software documentation and its date, communities available at its size, age of technology, popularity of technology. The factor 'Legal Restrictions' handles restrictions caused by licensing, which can be a barrier to software developers adding support for the format. This can be problematic when selecting an emulation strategy for long term preservation. The PREMIS metadata standard has semantic units for capturing this, that might need to be extended. The EU project 'KEEP' has a lot of case studies on this topic. This factor is dependent on licence and number of patents. The factor 'Lifetime' is characterised not only by the release date e.g. TIFF 6.0 was released in 1992. If a file format has had a long life, it has more support and it is less likely to become obsolete. This could make it more ubiquitous (e.g. JPEG or GIF). It can be considered associated to maturity and stability, or with outdated technology. Lifetime is dependent on release date, how many years its been supported by various well-known applications including web browsers, number of software, lifetime of supporting software. The factor 'Metadata Support' should support the file format longevity. The preservation institution can associate metadata with a file either by assigning of a persistent URI to the file and make sure that metadata contains the link, or using the base64 encoding in XML wrapper. The metadata offers support for validation and identification, content deduplication, indexing and retrieval. Some formats are very capable of embedding metadata (e.g. TIFF, WAVE). There

is a trend in new format specifications to include the ability to capture metadata inside of the object. If metadata is held externally, it allows knowledge to know about the file without having to open it and examine the internal metadata structures. Metadata attributes are documentation, characteristics of the file format, technical and descriptive metadata. The factor 'Rendering Software Available' is important for understanding when renderability is compromised and then institute the appropriate preservation planning, strategies and actions necessary to ensure it. This factor can be evaluated by testing, licencing, contacting vendors, using characterisation software and technology watch. The factor 'Revision Rate' is dependent on organization producing the software. This can be an indication of a volatile format as compared to a well-understood stable format. It is not very important if the 'old' versions will be still available. The best guarantee for format sustainability is multiple open-source renderers combined with a documented specification. Revision Rate attributes are number of years between revisions, average length of time for each revision, version control, defining terms, number of versions, documentation, number of rendering software. The factor 'Standardization' makes use of standards (e.g. ISO, or W3C) and increases the probability of the well formed specification. Some institutions may define local fields. Standards increase interoperability, popularity and migration support. The TIFF and WAVE lacks on standardization but nevertheless are widely adopted by preservation community. A format may still have patent-controlled elements. Standardization is dependent on specificity of the standard, specification, rendering/backward compatibility, conformance tests, breadth of adoption and complexity. The factor 'Storage Space' is not very relevant for format endangerment besides cost issue. Within a particular genre formats don't vary in size greatly. It depends if an organization can take advantage of compression or not, it also depends if a format is highly optimized for a particular content type. It is more important that files are stored multiple times to prevent and/or repair bit rot. Storage Space attributes are average file size and compression. The factor 'Technical Dependencies' addresses issue that the

more dependent a format is on a particular technological frame, the greater the necessity of maintaining some or all of that frame to insure continuing access. The required dependencies and technical environment should be captured as metadata. Technical Dependencies attributes are disclosure, IP issues, adoption, component pieces, dependency tree, feasible substitutions, documentation, emulation, metadata, OS, number of software and popularity. The factor 'Technical Protection Mechanism' (TPM) describes the danger of loosing access key by long time usage of format. This feature should be discouraged if the format supports it (e.g. PDF) and should be mentioned in metadata. The factor 'Ubiquity' takes care that data remains accessible. The assumption is that widely used format will be less likely subject to obsolescence. Depends however also on things like the viability of the supplier, whether it is proprietary or not and the emergence of new more interesting formats. Well used file formats have both active user communities and are more attractive to commercial companies to provide new products to support old formats. The more ubiquitous a file format is the wider is availability of toolsets for rendering, validation, identification, migration and emulation. Ubiquity attributes are market survey research, popularity, vendor information, proprietary, number of files, web search, number of software. The factor 'Value' determines that if there is enough valuable content in this format than its possible that someone will write software to support it. The value of information may not always be known when a format is chosen. Value attributes are specification, how hard/expensive is it to render the file compared to the value of the content, value known. The factor 'Geographic Spread' describes the level or breadth of adoption. If the usage is geographically spread, presumably many more companies and software around the world support it and the format is less vulnerable. Geographic Spread attributes are popularity, language, number of countries, number of users, number of vendors, number of people. The factor 'Domain Specificity' address the issue that some of the formats may be bind of a domain. Its attributes are number of domains and number of communities. The factor 'Viruses' is rather an attribute of the systems in which a format is encountered or managed and can be managed by migration. The factor 'Availability Online' can be regarded as the degree to which this format can be opened or read in a browser that emphasizes its popularity. It is dependent on number of software, support in browsers and documentation. The factor 'Institutional Policies' can be regarded as another dimension of adoption. The policies can change in time, and they cannot prevent the development of alternative file formats. The attributes of this factor is a consideration of risks, mitigating factors and resources. The factor 'Specification Quality' express expectation to a specification to be complete and well written. The better the specification the better any new implementation will be. As OAIS notes, sometimes source code for a renderer is itself representation information for a format. It is dependent on levels of satisfaction and specification.

An overview of the computed low level risks for the formats included in the evaluation set is presented in Table 2. The values and the interpretations of the most important 23 risk factors are presented. Within this representation, the "+" sign stands for *true* while the "-" sign means *false*. *L* depicts low risk, *M* means middle risk and *H* stands for high

Table 3: Exemplarily selected file formats with retrieved correctness information

Format	Expert Knowledge	Inaccuracies	Conflicts	Confidence Level
PDF	High	2	3	Middle
JP2	High	3	6	Low
JPG	Middle	1	1	High
JFX	Low	1	1	High
PNG	Middle	1	1	High
GIF	Middle	1	2	Middle
DOCX	Low	0	1	High
TIFF	High	0	1	High

risk. This table shows that among evaluated formats, the *DOC* format has the highest number of supported software, whereas for *SXW* only one software tool was documented in LOD repositories. The remaining formats have different software numbers, mostly between 10 and 40.

The different risk scores for *DOC* (low) and *PPT* (middle) could be explained with larger amount on software tools automatically detected for *DOC* (164) comparing to four for *PPT* and also with more descriptions for *DOC* format. Additionally, for *DOC* the genre, creation date, publisher and creator information were retrieved, whereas these factors are missing for *PPT*. This does not mean that such information does not exist for *PPT*, it only indicates that this is not included or not found in LOD repositories. The same consideration is valid for the "software count" value 12 of *MP3* format. It is known that there should be much more associated software tools that are able to handle this format.

At this point it should be stated that not all formats were analyzed and that evaluated results currently require verification by human expert and further optimisation of calculation methods. Evaluation results presented in Table 2 are limited to the information automatically collected from mentioned above LOD repositories and is customized by applied expert rules. Therefore these results cannot be regarded as absolutely accurate, but they provide a good overview of the possible preservation risks related to the given file formats. The classification settings for risk factors are institutional dependent and is a matter of discussion and a future work. The default thresholds are defined based on the accessible expert knowledge and could be customized according to preferences of particular user.

4.3 Detailed Report

The evaluation demonstrates 3 that given approach shares expertise and supports contradictions comparison for one institution and addresses specific risks within file formats. Information support provided by the Expert System helps in solving practical digital preservation issues. But in order to generate higher value in aggregating the data sources and exposing conflicts and inaccuracies this tool needs more and better quality data sources.

Although FFMA provides valuable information that well describes the evaluated formats, the accuracy of data collected in FFMA knowledge base should be examined by experts. The PDF is marked as not compressed format, but experts state that PDF nearly always uses flate compression, whereas a whole array of compression methods may be used for images. PNG, JPG and GIF are flagged in FFMA as uncompressed whereas they have compression. The Jpeg2000 format according to FFMA is not supported by any soft-

Table 2: Exemplarily selected file formats with retrieved information for associated measurement metrics

Risk Factor	GIF	PNG	MP3	PDF	JPG	DOC	HTML	TIF	BMP	PPT	MAC	SXW	DXF
Is Popular Format	5/L	5/L	5/L	5/L	5/L	5/L	5/L	5/L	5/L	5/L	2/H	3/M	5/L
Operation Systems	3/M	4/L	3/M	6/L	4/L	5/L	4/L	3/L	2/M	5/L	2/M	3/M	4/M
Software Count	18/M	21/M	14/M	28/M	17/M	164/L	39/L	135/L	18/M	15/M	122/L	1/H	21/M
Vendors Count	3/L	1/M	3/L	2/L	1/M	1/M	1/M	1/M	1/M	1/M	1/M	1/M	1/M
Versions Count	2/M	3/M	1/L	17/H	9/H	15/H	7/H	9/H	7/H	7/H	1/L	1/L	23/H
Has Description	3/M	3/M	2/H	3/M	2/H	3/M	2/H	3/M	2/H	2/H	2/H	2/H	2/H
Has MIME type	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H
Existence Period	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Is Complex Format	-/L	-/L	-/L	+/H	-/L	-/L	+/H	+/H	-/L	-/L	-/L	+/H	+/H
Is Wide Disseminated	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H
Is Outdated or Deprecated	-/L	-/L	-/L	-/L	-/L	+/H	+/H	-/L	-/L	+/H	+/H	+/H	+/H
Has Genre	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H	-/H	-/H
Has Homepage	+/L	-/H	-/H	+/L	-/H	-/H	-/H	-/H	+/L	-/H	-/H	-/H	-/H
Is Open (Standardised)	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H	-/H	-/H
Has Creation Date	+/L	+/L	+/L	+/L	-/H	+/L	+/L	+/L	-/H	-/H	-/H	-/H	-/H
Has File Migration Support	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Digital Rights Information	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H
Has Publisher Information	+/L	-/H	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H
Has Creator Information	+/L	-/H	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H
Has Compression Support	-/L	-/L	-/L	-/L	-/L	-/L	-/L	+/H	-/L	-/L	-/L	-/L	-/L
Supported by Web Browser	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Has Vendor Support	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L

ware and does not have a MIME type, is frequently used and is supported by web browsers. In reality these factors are wrong in FFMA. The JPX format is marked as not a compressed that should be less complex than JP2, but actually it is an extension of Jpeg2000 with added complexity. The GIF is marked as having the highest risk. The TIFF format should have higher risk than PDF or DOCX. The PDF can be a container for Jpeg2000 which is considered high-risk in FFMA. The mentioned confidence levels should not be regarded as a preservation risk estimation for associated format. Currently FFMA provides generalized information about formats, without addressing specific risks within formats. It should be mentioned that presented confidence levels are considered in relation of FFMA results to expert knowledge. These are FFMA evaluation results and should help the user to resolve these contradictions.

5. CONCLUSIONS

In this work we presented an approach for bringing together information automatically aggregated from open sources and an expert knowledge related to digital preservation. The main contribution of this work is the definition and computation of fuzzy logic for metrics generation in order to support digital preservation experts in semi-automatic estimation of “institutional obsolescence” for file formats. We aggregated a solid knowledge base from linked open data repositories. In correctness report we expose conflicts and inaccuracies in these data in order to improve the quality of a risk analysis in the digital preservation domain. This method facilitates decision making with regard to preservation of digital content in libraries and archives using expert knowledge as a basis. We have developed a tool for aggregating file format descriptions that exploits available linked data resources and uses expert models to infer knowledge regarding the long-term preservation of digital content. The ontology mapping technique that comprises expert rules and clustering is employed for collecting the information from the web and integrating it in a common representation. We employed fuzzy logic techniques for processing of aggregated information about formats in form of metrics in order to bring conflicted and incorrect information to the surface for correction and improvement by community. The analysis results of a survey regarding the risk factors for file formats was integrated in a fuzzy model and is presented in evaluation part. The evalua-

tion demonstrates that given approach shares expertise and supports contradictions comparison for one institution and addresses specific risks within file formats. Information support provided by the Expert System helps in solving practical digital preservation issues. But in order to generate higher value in aggregating the data sources and exposing conflicts and inaccuracies this tool needs more and better quality data sources. The analysis and measurement provided by developed Expert System is about the reduction of uncertainty and not about the elimination of it. Using our system with its metrics we have ability to measure and the ability to think about how could we use these measurements.

As future work we plan to increase the amount of an aggregated information, to extend an Expert System with additional fuzzy rules and to improve its accuracy and quality of the outputs.

6. ACKNOWLEDGMENTS

This work was partially supported by the EU FP7 Project SCAPE (GA#270137) www.scape-project.eu. The authors wish to thank Paul Wheatley from the British Library for his thoughts on the topic and Ryan Heather for providing of the results from the endangerment study on file formats.

7. 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] 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.
- [4] 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.

- [5] 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.
- [6] R. Heather. File format study. *School of Information and Library Science, University of North Carolina at Chapel Hill*, 2, 2013.
- [7] 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.
- [8] 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.
- [9] C.-C. Lee. Fuzzy logic in control systems: fuzzy logic controller. i. *Systems, Man and Cybernetics, IEEE Transactions on*, 20(2):404–418, 1990.
- [10] 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.
- [11] D. S. Rosenthal. Format obsolescence: assessing the threat and the defenses. *Library Hi Tech*, 28(2):195–210, 2010.
- [12] J. Rothenberg. Digital preservation in perspective: How far have we come, and what’s next? *Future Perfect 2012*, 2012.
- [13] H. S. Sii, T. Ruxton, and J. Wang. A fuzzy-logic-based approach to qualitative safety modelling for marine systems. *Reliability Engineering & System Safety*, 73(1):19 – 34, 2001.
- [14] S. Vermaaten, B. Lavoie, and P. Caplan. Identifying threats to successful digital preservation: the spot model risk assessment. *D-Lib Magazine*, 18(9/10), September 2012.
- [15] L. A. Zadeh. Fuzzy logic = computing with words. *Fuzzy Systems, IEEE Transactions on*, 4(2):103–111, 1996.