# A Risk Analysis of File Formats for Preservation Planning[*]

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 automatic estimation of preservation risk for file formats. The main contribution of this work is a definition of the risk factors with associated severity level and its automatic computation. Our goal is to apply a solid knowledge base automatically extracted from linked open data repositories as the basis of the risk analysis system for digital preservation. This method is meant to facilitate decision making with regard to preservation of digital content in libraries and archives. The File Format Metadata Aggregator tool is employed in order to aggregate well founded and trusted file format information through linked data and inferred knowledge in the domain of long-term information preservation. The ontology mapping technique is employed for collecting the information from the web of linked data and integrating it in a common representation. Furthermore, we employ AI technologies (i.e. expert rules, clustering) for inferring explicit knowledge on the nature and preservation friendliness of the file formats. A statistical analysis of the aggregated information and the qualitative analysis of the aggregated knowledge are presented in the evaluation part of the paper. A Web service is created to support programmatic access to format and risk analysis reports.

## 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

The core of preservation planning aims at creating of sustainable storage solution for digital collections are file formats using for encoding the digital information. Currently the information about the file formats is not structured or is only partly structured. The preservation risks for particular file format are difficult to estimate and definition of risk factors is unclear. Intensive human expert involvement is required for search, aggregation and estimation of format risk information. The definition of risk factors for preservation risk can vary depending on preservation goals and preservation workflows of particular organisation. Also classification and weighting of risk factors is a challenging task and is strongly dependent from an expert knowledge and experience. The existing domain specific knowledge bases do not contain all necessary semantic information about file formats or their description fields or such information is not sufficient. The list of the maintained file formats is not complete and differs from source to source. The richness of knowledge base information is important to support decisions on preservation planning regarding risk estimation. Even though the world wide web has turned out to be the largest knowledge base, the information published lacks an unified well-formed representation. The linked open data (LOD)[1] and Open Knowledge[2] initiatives address these weaknesses by describing a method on how to publish structured data in a well-defined and queriable format. In order to aggregate sufficient knowledge about file formats for risk analysis we linked together different independent and publicly available information sources like Freebase[3], DBPedia[4] and PRONOM[5].

The PRONOM registry provides persistent, unique, and unambiguous identifiers for file file formats and therefore takes a fundamental role in the process of managing electronic records. The most of the file formats are properly documented, are open-source and well supported by producer. The others formats may be outdated, redeemed by software vendors and no longer functional with modern software or hardware. Some customized file formats could be obsolete and not accessible. To get a grip on all these problems we use the File Format Metadata Aggregator (FFMA) ([5]) system depicted in Figure 1, which aims of preparing the ground for KBRs like DiPRec [4]. The FFMA reuses the experi-

[*] (Does NOT produce the permission block, copyright information nor page numbering). For use with ACM_PROC_ARTICLE-SP.CLS. Supported by ACM.

[1] http://linkeddata.org/

[2] http://www.okfn.org/

[3] http://www.freebase.com

[4] http://dbpedia.org/

[5] http://www.nationalarchives.gov.uk/PRONOM/

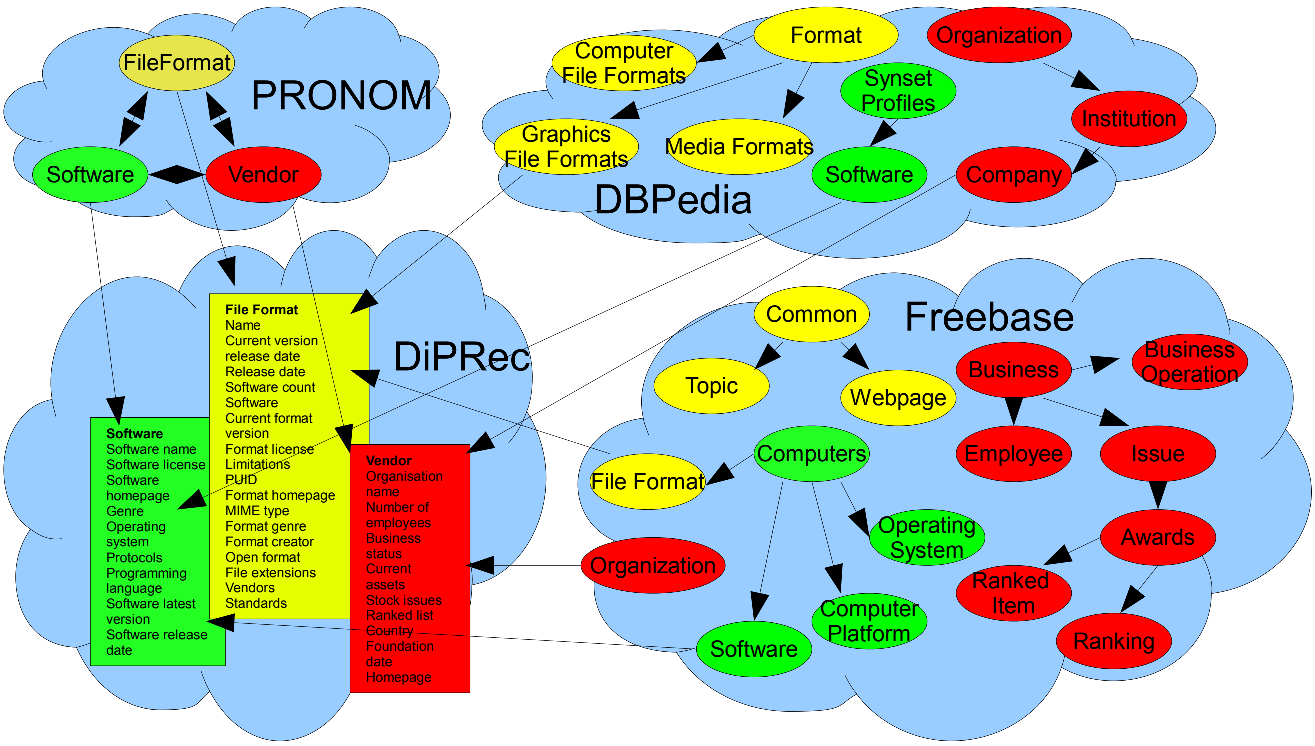


**Figure 1: PRONOM, DBPedia and Freebase digital preservation domain related ontology sections mapped to the DiPRec file format ontology.**

ence of building preservation planning tools and addresses the topic of digital long-term preservation and provides data analysis based on the concept of risk scores. The knowledge base is built through a LOD approach. Information regarding file formats and software vendors is taken into account and retrieved from Freebase, DBPedia and PRONOM. The important contribution of this paper consists in the technical information analysis and assessment regarding preservation risks for different formats. Another contribution is using of LOD ontologies mapping (Figure 1) for risk analysis and in the integration of linked data when designing the knowledge base. Decision support based on elaborated rule engine provided by FFMA is meant to support institutions like libraries and archives with suggestions in the process of analyzing their digital assets. FFMA collects and structures format information following by semi-automatic open datasets extraction from the linked data repositories independent from provenance of the data registry and query language. We aim at designing well founded knowledge base with defined rules and scored metrics that could be used for decision making support for quality assurance of document image collections. The paper is structured as follows: Section 2 gives an overview on related work and concepts. Section 3 explains knowledge base aggregation process and covers also ontology mapping, rule engine and algorithmic details of risk analysis. Section 4 presents the experimental setup, applied methods, description of the web service for risk analysis and results. Section 5 concludes the paper and gives outlook about planned future work.

## 2. RELATED WORK

In [7] Andrew Jackson tested competing hypotheses regarding software obsolescence issue employing format identification tools in order to select preservation strategy. One of these hypothesis is presented by Rothenberg [13] and means that all formats should be considered brittle and transient, and that frequent preservation actions will be required in order to to keep data usable. In contrast to that hypothesis the Rosenthal [12] claims that no one supporter of format migration strategy was able to identify even one format that has gone obsolete in the intervening decade and a half. Rosenthal argues that the network effects of data sharing inhibit obsolescence. But 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. In Andrew Jackson research he examines how the network effects could stabilise formats against obsolescence in order to understand the warning signs, choices and costs involved. This evaluation should help to meet preservation strategy: either to perform frequent preservation actions to keep data usable 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 goal of the SPOT (Simple Property-Oriented Threat) model [14] is to help repositories identify previously unaddressed threats, perform preservation risk monitoring, and demonstrate the repository compliance to the accepted standards. In this work the digital preservation risks are divided into two categories: threats to archived digital content, and threats to the custodial organization itself. The SPOT Model focuses on the first category and develops a framework for assessing threats arising from the technical operations associated with preserving digital objects. The SPOT risk model is limited to properties like availability, identity, persistence, renderability, understandability and authenticity. But these properties do not define measurable risk factors and do not employ LOD repositories information.

The AONS II tool [11] aims at recognising file formats in a digital collection, retrieving of information on obsolescence risk indicators by reference to external registries and building collection profiles. This tool is able to distinguish accurately between different versions of formats, in order to identify relevant risk levels. AONS II tool struggles to solve problems like misleading file extensions and different names for the same format by creating of internal format identifier for each apparent format found, and then tries to map it to the likely matching format identifiers used by external registries. But this tool does not apply risk factor metrics for risk calculation. Reading the [11] about the AONS tool we realized the need to develop a central web service in order to be able to share the results of local risk assessments. Creating this service we aim at defining of risk metrics based on experience of community members which share their individual risk findings. This would allow LOD registries to leverage the experiences and expertise of the contributing preservation community and add considerably to their usefulness.

In our approach we are not intend to mark format as an obsolete format, since there are different hypotheses about format obsolescence and we do not consider if a format is likely to be obsolete or not in a binary fashion. We define obsolescence in relation to how much extra it will cost to render a file beyond the capability of a standard vanilla PC setup in particular institution. In our approach we analyze a format regarding it's "institutionally obsolescence" (term introduced by Paul Wheatley). That means that a particular format would not render on a PC in an institution's reading room. With our system we aim at assessing the risks associated with format rendering. We use the risk factors like "is compressed", "is supported by web browser", "has supporting software", "has supporting vendor", "versions count", "creator information", "operation system", "genre", "creation information", "is migration supported", "has digital rights information" etc. Many of these factors have influence on rendering. The advantage of our tool is that it enables configuration of risk factors by user and he could use it according with his convincements.

The format risk analysis approach in [3] presents an P2 registry, which is an RDF-based framework. P2 registry employs information containing in DBPedia and Pronom repositories and supports its own format risk analysis system. The main goal of P2 platform to allow and encourage publication of preservation data. This repository calls for the active participation of the digital preservation community to contribute data by simply publishing it openly on the Web as linked data. In contrast to the P2 registry the FFMA tool does not use RDF triples and additionally to DBPedia and Pronom makes use of the rich Freebase repository. FFMA tool has capability to extend repositories pool to any other repository independent from repository architecture. In our approach we present the rule engine for risk analysis that handles risk factors not covered by P2 registry and supports definition of customized risk factors. The advantage of these risk analysis model is that model is customizable by changing severity weights of particular rules according to user requirements. Additional user specific rules can be simply added to the model.

The PANIC tool [6] was aiming at automatically inform repository managers of changes that might cause risks for accessibility of their collections and alerting when file formats become obsolete. The idea of this tool is to aggregate data and metadata for further analysis, but this information is not easy accessible and collected information could be more representative. Also there is no common understanding in the community about the meaning of term "obsolete" as mentioned above.

Existing tools for long term preservation planning like Plato ([9, 1]) enable different digital preservation actions like identification, characterization and migration. These tools present information about possible preservation action but do not provide suggestions or recommendations regarding format preservation risks for user that do not have an expertise in the digital preservation domain.

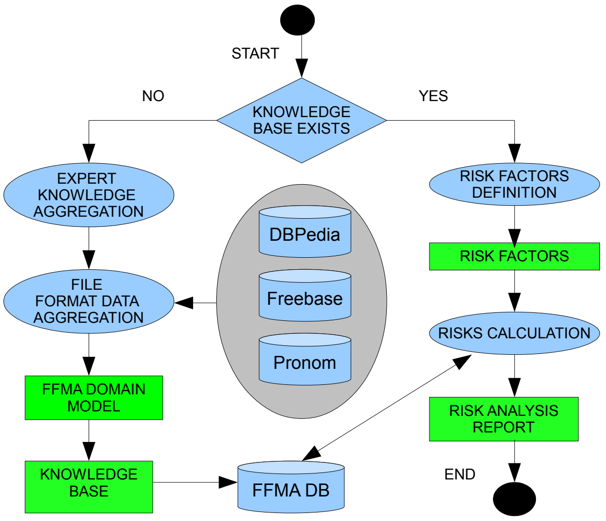


**Figure 2: The format risk analysis workflow.**

## 3. SYSTEM OVERVIEW

The FFMA system overview is presented in Figure 2 by the risk analysis workflow. The basis of risk computation is a properly created knowledge base that stores its data in FFMA database. An evaluation of expert knowledge and an aggregation of the file format data are necessary steps in order to build the knowledge base. Having FFMA database at hand we define and employ risk factors for computation of file format risks. The results of risks calculations are presented in reports provided in HTML format. In the case that FFMA database already exists the risk analysis workflow starts directly with risk calculation.

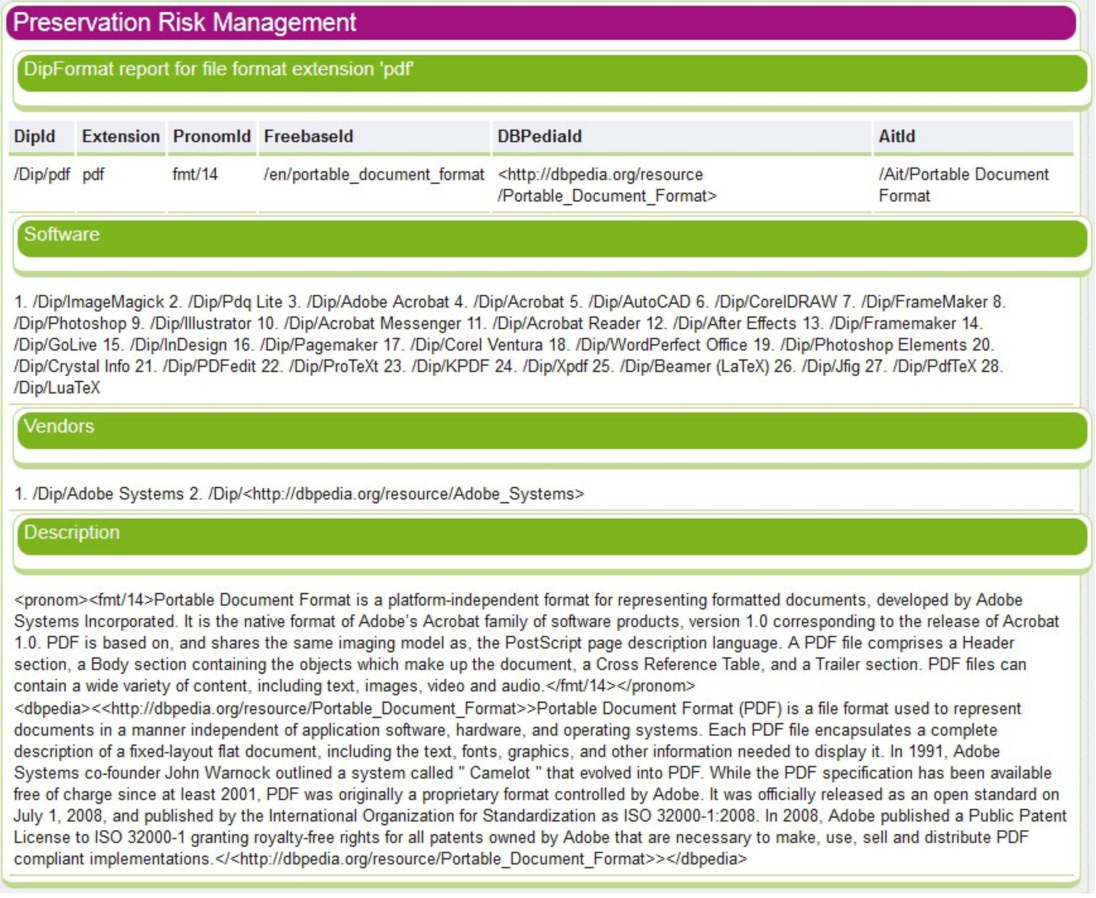
## 3.1 Aggregation of File Format Data

The FFMA module for aggregation of file format data collects information on file formats and enhances it by module for expert knowledge aggregation therefore building the knowledge base. At runtime takes place data analysis on knowledge base implying the underlying FFMA domain object model and based on the user's input configurations. These configurations specify which LOD repositories should be used and which file format properties are of interest for particular institution. 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 after ontologies mapping in Expert Knowledge Aggregation module. Scalable information extraction is provided by storing of domain knowledge in a database. We consider Freebase [10] as one of the most valuable sources for information extraction. It is a practical, scalable semantic database for structured knowledge. The PRONOM ontology looks very similar to the FFMA ontology but doesn't contain all necessary properties (like genre or vendor business status) that DiPRec requires to incorporate significant data from another ontologies. Internal mapping between different knowledge bases was set up to enable data assignment because e.g. multiple vendor definitions for the same vendor are possible. Extending the PRONOM repository, we collect additional information sources and aggregate them in a single homogeneous property representation in the FFMA knowledge base. The data representation retrieved from different knowledge bases is unified and combined in FFMA domain model. The individual object's namespace, the transformation process of values, the query on how to extract a given record, etc. are preserved and are part of the property's model representation. When aggregating file format data we are employing external knowledge sources like DBPedia and Freebase which manage huge amounts of LOD triples. This allows us to extract fragmental descriptions on file formats, software applications, and vendors supporting given file formats. DBPedia allows to post sophisticated queries using SPARQL query and OWL ontology languages [8] for retrieving data available in Wikipedia. Public read/write access to Freebase is allowed through an graph-based query API using the Metaweb Query Language (MQL) [2]. PRONOM data is released as LOD and is accessible through a public SPARQL endpoint.

The list of the file formats automatically extracted from selected LOD repositories is a result of the process of file format data aggregation and is then used in conjunction with domain knowledge as a knowledge base and an input for FFMA engine for decision support. File format properties are designed to give an option at hand for definition of user rules, metrics and classifications. File format property object comprises also queries to LOD repositories, classification objects, definitions of risk factors and property descriptions. The risk factors are used to compute recommendations based on user input. The experimental knowledge of data registries helps us to establish methodology and gives some rough estimates about which risk properties should be defined with which risk classification. The most significant data repository queries in terms of digital preservation addresses PUID, file formats, software and software vendors. Relationships between these parameters and accompanying parameters like computer platform, genre, license, programming language, release date, homepage, compression type and so on are of interest for risk analysis. Optionally user is able to extend default risk analysis model adding his own property sets and classifications using correspondent configuration files. In order to reduce the required domain knowledge acquisition efforts the knowledge base stores the aggregated information in FFMA domain object model. After initial storage we only need to update particular database areas. This model increases performance. The potential drawbacks during the database initialization could be e.g. queries limit, bad internet connection to repositories or server could be offline for maintenance purposes.

## 3.2 Risk Factors Definition

In order to organize the Knowledge Base we must structure the information that has been obtained from the domain experts of digital preservation and from conducted experiments. We define typical scenarios and identify the parameters used by library experts for collection handling. Then we define the linguistic labels to classify measured values of each parameter and associated ranges. Finally, we determine the conditional rules that relate these linguistic labels to specific consequences. The knowledge acquisition for the Knowledge Base is performed by librarians who provide the knowledge engineer with typical application use cases, metrics and parameters that characterize the preservation processes. Information retrieved from the image collection is processed by the customized domain model. This model enables structured and maintainable handling of analyzed data. If necessary, the data could be stored in a database for further treatment. A user communicates with the Expert System by sending a request query and receives an advice in response. The most significant risk factors describe how many software tools and vendors support particular file format. The version count metric could be interpreted in different ways. On the one hand the more versions format has the more work is invested in its development and support. And that means this format is in use and good supported. On the other hand the more versions has a particular format the higher is the probability that it will cause confusing in digital preservation workflows e.g. for migration process. Changing severity value and classification settings each customer could adjust the meaning of this risk factor for his specific needs and understanding. Documentation level is also an important risk factor. Additional help for risk estimation provide specification factors like whether a format has a homepage, genre definition, creator and publisher information, is supported by web browsers, has compression. Nowadays very important issue become digital rights information. For preservation processes it is important to know whether format migration is supported. The MIME type provides a connection chain between different repositories. The complexity of the file format could be measured by assessment of documentation, format standard, relation



Preservation Risk Management

DipFormat report for file format extension 'pdf'

| DipId | Extension | PronomId | FreebaseId | DBPediaId | AitId |
|---|---|---|---|---|---|
| /Dip/pdf | pdf | fmt/14 | /en/portable_document_format | <http://dbpedia.org/resource /Portable_Document_Format> | /Ait/Portable Document Format |

Software

1. /Dip/ImageMagick 2. /Dip/Pdq Lite 3. /Dip/Adobe Acrobat 4. /Dip/Acrobat 5. /Dip/AutoCAD 6. /Dip/CorelDRAW 7. /Dip/FrameMaker 8. /Dip/Photoshop 9. /Dip/Illustrator 10. /Dip/Acrobat Messenger 11. /Dip/Acrobat Reader 12. /Dip/After Effects 13. /Dip/Framemaker 14. /Dip/GoLive 15. /Dip/InDesign 16. /Dip/Pagemaker 17. /Dip/Corel Ventura 18. /Dip/WordPerfect Office 19. /Dip/Photoshop Elements 20. /Dip/Crystal Info 21. /Dip/PDFedit 22. /Dip/ProTeXt 23. /Dip/KPDF 24. /Dip/Xpdf 25. /Dip/Beamer (LaTeX) 26. /Dip/Jfig 27. /Dip/PdfTeX 28. /Dip/LuaTeX

Vendors

1. /Dip/Adobe Systems 2. /Dip/<http://dbpedia.org/resource/Adobe_Systems>

Description

<pronom><fmt/14>Portable Document Format is a platform-independent format for representing formatted documents, developed by Adobe Systems Incorporated. It is the native format of Adobe's Acrobat family of software products, version 1.0 corresponding to the release of Acrobat 1.0. PDF is based on, and shares the same imaging model as, the PostScript page description language. A PDF file comprises a Header section, a Body section containing the objects which make up the document, a Cross Reference Table, and a Trailer section. PDF files can contain a wide variety of content, including text, images, video and audio.</fmt/14></pronom> <dbpedia><<http://dbpedia.org/resource/Portable_Document_Format>>Portable Document Format (PDF) is a file format used to represent documents in a manner independent of application software, hardware, and operating systems. Each PDF file encapsulates a complete description of a fixed-layout flat document, including the text, fonts, graphics, and other information needed to display it. In 1991, Adobe Systems co-founder John Warnock outlined a system called " Camelot " that evolved into PDF. While the PDF specification has been available free of charge since at least 2001, PDF was originally a proprietary format controlled by Adobe. It was officially released as an open standard on July 1, 2008, and published by the International Organization for Standardization as ISO 32000-1:2008. In 2008, Adobe published a Public Patent License to ISO 32000-1 granting royalty-free rights for all patents owned by Adobe that are necessary to make, use, sell and distribute PDF compliant implementations.</<http://dbpedia.org/resource/Portable_Document_Format>></dbpedia>

**Figure 3: Sample aggregated data report for PDF file format.**

between different versions of the same format, compression factor etc. Some formats are implicitly or explicitly declared as outdated or deprecated. Format standardisation reduces the risk of its using. The existence period of a format is an additional metric for risk estimation, since the longer and popular is a particular format the lower is the risk. Software, vendors and versions count factors together with description factor build an aggregated rule whether the given format is supported by FFMA. Missing of these important information means that regarded LOD repositories does not provide information about required format. The previously defined rules should be organized in order to process input statements (assertions) and to infer appropriate advice and conclusions. Forward rule chaining for file format analysis is presented in Figure 4. Forward chaining is the process of moving from the "if" patterns (antecedents) to the "then" patterns (consequents) in a rule-based system. We consider the antecedent as satisfied when the "if" pattern matches the assertion. Assertions are depicted by black rectangles on the input side and by the white rectangles on the output side, respectively. The rules are presented by blue half-spheres. A specific rule is triggered if all of its antecedents are satisfied. A triggered rule is considered as fired if it produces a new assertion or performs an action on the output (white rectangle). Since our rule engine is focused on file format analysis there is no need for any conflict-resolution procedure to resolving possible rule conflicts. In Figure 4 we present rules distinguishing low level risk format from high level risks. The rule-base system starts risk identification with the rule D1. Suppose that software count is higher than 0. Then if the antecedent pattern defined in classification settings matches that assertion, the value x becomes "is supported by software" and the rule D1 fires. Because the aggregated risk of rules D2, D3 and D4 matches the antecedent patterns for vendors, versions and descriptions count and has acceptable risk level severity, rule D22 fires, establishing that the format exists in aggregated knowledge base. This fact enables further analysis and we could similarly go through remaining rules. The final conclusion of the rule-based system is whether an analysed file format has high, middle or low preservation risk and which particular risk factors cause this risk.

## 3.3 Risk Computation

The evaluation of the risk score presented in Figure 4 can be conducted when the risk analysis task fills previously created risk analysis model with risk factor values from different knowledge bases. Each risk value is selected from associ-

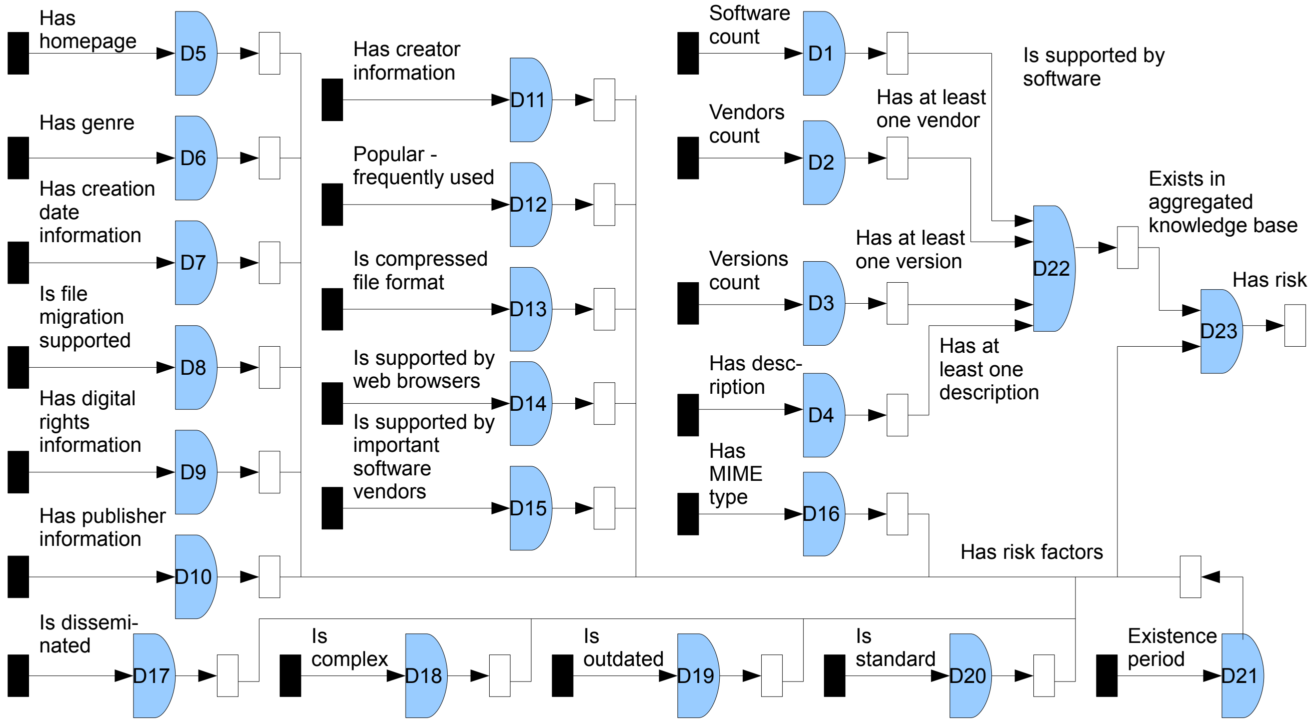


**Figure 4: Forward rule chaining for risk analysis.**

ated risk factor depending on response of the knowledge base for correspondent request query defined in a risk property. The risk score for particular property is then evaluated from risk analysis model dependent on metrics, property weight and risk classifications. The report on risk analysis provides contributed content inspection to categorize it based on its preservation risk. The risk report contains information about computed risk properties, associated risk factors and evaluated risk values inserted in risk analysis model at run time as a result of calculations for the risk analysis. Due to management and maintenance reasons properties are grouped by sets. A property may belong to one or more property sets. The extent to which a property belongs to a property set and consequently contributes to the risk computation over a given dimension is modeled through the introduction of specific weighting factors (see Equation 1). The computation of the overall risk score for FFMA properties is presented in [4] and is computed as a weighted sum over all risk factors:

$$R_i = \sum_{ps \in PS_i} w_{ps,i} * \sum_{p \in PROP_{ps}} w_{p,ps} * d(p, PFV(p)) \qquad (1)$$

Where $R_i$ represents the preservation risk computed over the preservation dimension $i$, $ps$ represents the index of the current property set within all sets associated to the dimension $i$ ($PS_i$). The $w_{(ps,i)}$ is the weight of the contribution of the property set $ps$ to dimension $i$. Similarly $p$ stands for the index of current properties within the list of properties available in the given property set $PROP_{ps}$. $w_{p,ps}$ denotes the importance of a property $p$ for the property set $ps$. The distance between the current property and the defined - 'preservation conform' - value for this property is represented through $d(p, PFV(p))$.

## 4. FORMAT RISKS EVALUATION

The evaluation of format risks was conducted with DiPRec repository based on FFMA database. Our hypothesis is that file format data automatically aggregated from LOD repositories will provide the rule engine with valuable information and will enable risk estimation for different file formats. We also expect that the distribution of calculated format risk scores will match to the associated information that we could find in the domain literature. The "low risk" marked formats should indicate the currently most reliable file formats for digital preservation workflows. One of the most important use cases for FFMA system is an evaluating of software solutions available for processing of the preservation plans and its assessment regarding preservation risk. We developed a Web service that automatically retrieves file format related data from LOD repositories and performs reasoning on collected information employing specified risk factors. The basis of this service relies on rich data descriptions retrieved from LOD repositories like DBPedia, Freebase, PRONOM etc. The LOD data is automatically harvested using the Web access points provided by these repositories and the different request supported query languages. The collected information is processed, normalized, integrated into the knowledge base of the service and subsequently classified in order to calculate risk scores for particular file format. The programming interface of this service supports quering 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. Service provides generation of rich format descriptions and a report on format risks.

## 4.1 Dataset Description

A sample of aggregated data report for PDF file format is generated as HTML table and is presented in Figure 3. This automatically created report comprises FFMA identifier "/Dip/pdf", unique identifiers of the repositories that contain information regarding given extension "pdf". In this case these are PRONOM with "fmt/14" and Freebase with "/en/portable_document_format" identifiers. Additionally we have got information about 28 different software tools and one vendor associated with this file format and presented by unique FFMA identifiers. Two LOD repositories provide different descriptions for the given file format. Since aggregated information is stored in a database, calculation time of the report demonstrates high performance (lower then a second). Aggregated reports on file formats contain information like "FileFormatDescription", "SoftwareName", "RepositoryName", "SoftwareHomepage", "SoftwareDescription" etc. These reports also include references to LOD repositories. According to the LOD principles, each linked data repository has its own mechanism for nonambiguous referentiation of the managed objects. By having a reference in a correct format, a user is able to easily request the information from a web service. As example, the references for "pdf" file extension look like following: in DBPedia [6] in Freebase [7] in PRONOM fmt/14 [8]. FFMA returns evaluated software, vendor and risk report objects in HTML format. The LOD domain objects support storage, retrieval, and analysis of information retrieved from LOD repositories. This structured information is a knowledge base to be used for deriving preservation recommendations.

For evaluation we analyse a subset of 13 well known file formats. The "GIF", "PNG", "JPG" and "BMP" represent the genre of image formats. The "TIF" is also a graphic format that additionally supports layers, multiple pages and compression. With "MP3" we have an audio format. The "PDF" format comprises multiple versions and is mostly used for documentation accompanied with Adobe Acrobat tool. The "HTML" format also has multiple versions and is used for creation of Web pages. The "DOC" and "PPT" are Microsoft formats for working with text 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 and supports only greyscale graphics. The "SXW" is an outdated text format for OpenOffice tool. The "DXF" is a vector graphic format for AutoCAD tool.

## 4.2 Application of rules and risk factors

Employing knowledge base in the way as described in chapter 3 we apply risk factors to the aggregated from LOD repositories information that is relevant to the file formats. In the Table 1 we demonstrate exemplarily selected file formats with semi-automatically retrieved information for associated risk factors. There we present 23 most important risk factors for 13 exemplarily selected file formats. Plus stands for "true" and minus means "false". Letter "L" depicts low risk, "M" means middle risk and "H" stands for the high risk. In our approach we regard different extensions of one format like "TIF" and "TIFF" as one format. User should not care about this issue since it is done automatically. This table demonstrates that among evaluated formats the "DOC" format has the highest number of supported software, whereas for "SXW" extension in LOD repositories was found only one software tool. The remaining formats have different software numbers, mostly between 10 and 40. Therefore we consider the risk regarding the "software count" risk factor for "SXW" file extension as high, risk for "DOC", "HTML", "TIF" and "MAC" extensions as low and medium risk is associated with remaining formats. Defining classifications for this risk factor we expect that the more software tools support particular file format the lower is its risk. But this factor can be also configured according to the idea, that many software tools could cause instability of file format. In this case user just need to redefine classification settings according to his risk estimations. The lowest risk for "vendors count" risk factor were calculated for "GIF", "MP3" and "PDF" formats with two to three vendors. The highest risk regarding "vendors count" risk factor was not detected at all. We expect high vendor risk in the case that no vendors were found for particular format. Remaining formats have middle risks for vendors. We assume that the more versions are defined for a format the higher is the probability of version confusion. Therefore our calculation evaluated the highest "versions count" factor risk for "DXF" (23), "PDF" (17), "DOC" (15), "JPG" (9), "TIF" (9), "HTML" (7), "BMP" (7) and "PPT" (7). The lowest risk for "versions count" factor demonstrate "MP3" (1), "MAC" (1) and "SXW" (1) formats. Remaining formats have middle risks regarding "versions count". Regarding descriptions we think that the more different descriptions were found the lower is the risk. According to this definition we have high risk for "MP3" (1), "JPG" (1), "HTML" (1), "BMP" (1), "PPT" (1), "MAC" (1), "SXW" (1), "DXF" (1) formats and middle description factor risk with values in range from two to three for remaining formats. All of the regarded formats have multiple descriptions but do not exceed threshold of three and therefore we don't have low risk among them. The MIME type is an essential reference in order to address a file format and to cre-

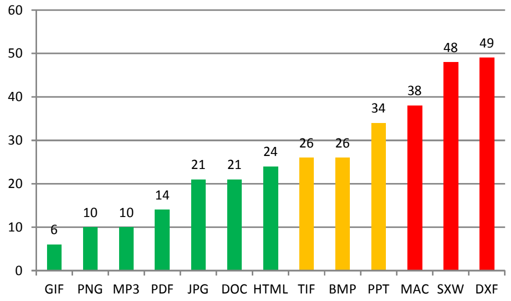


**Figure 5: The distribution of the file formats with associated risk scores in range from 0 to 100 percent. Where 0 stands for the lowest possible risk and 100 for the highest format risk.**

[6] http://dbpedia.org/resource/Portable_Document_Format

[7] http://www.freebase.com/view/en/portable_document_format

[8] http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=613&strPageToDisplay=summary

Table 1: Exemplarily selected file formats with retrieved information for associated risk factors

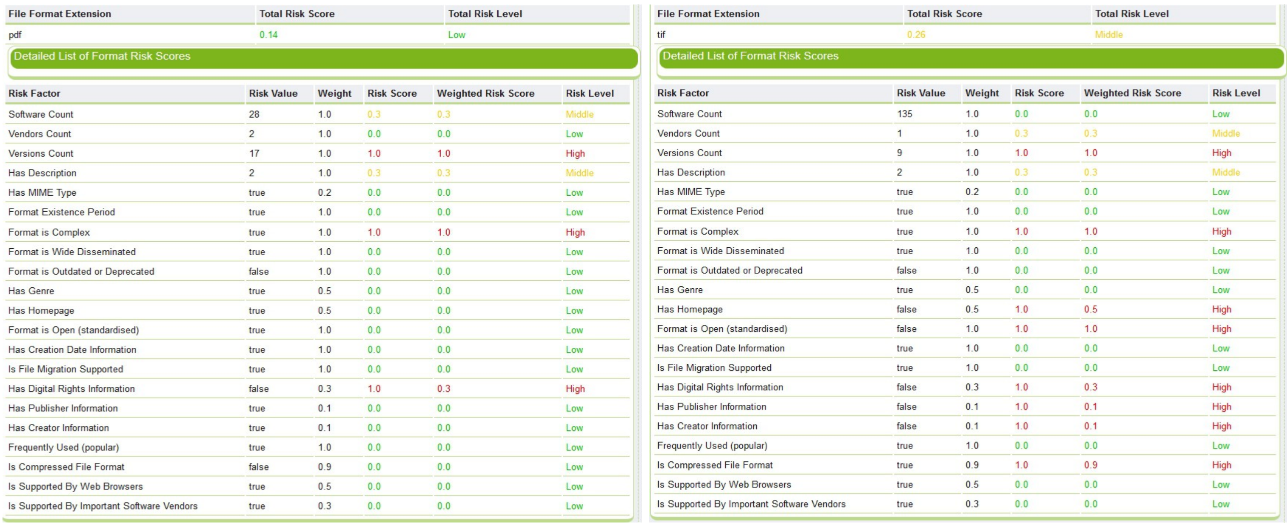
| Risk Factor | GIF | PNG | MP3 | PDF | JPG | DOC | HTML | TIF | BMP | PPT | MAC | SXW | DXF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Software Count | 18/M | 21/M | 12/M | 28/M | 17/M | 164/L | 39/L | 135/L | 18/M | 4/M | 122/L | 1/H | 9/M |
| Vendors Count | 3/L | 1/M | 3/L | 2/L | 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 | 2/M | 2/M | 1/H | 2/M | 1/H | 2/M | 1/H | 2/M | 1/H | 1/H | 1/H | 1/H | 1/H |
| Has MIME type | +/L | +/L | +/L | +/L | +/L | +/L | +/L | +/L | +/L | +/L | -/H | -/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 | -/H | -/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 | -/H | -/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 |
| Is Popular Format | +/L | +/L | +/L | +/L | +/L | +/L | +/L | +/L | +/L | +/L | -/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 |
| Total Risk (%) | 6/L | 10/L | 10/L | 14/L | 21/L | 21/L | 24/L | 26/M | 26/M | 34/M | 38/H | 48/H | 49/H |

ate a connection between different file format ontologies or identification tools. Most of presented formats have found associated reference. Only three formats are missing the MIME type. These are “MAC”, “SXW” and “DXF” formats. The longevity of the format existence period could give as a rough estimation about its stability. Therefore the longer a format is in use the lower is the preservation risk. In our case all of the formats have low risk in this regard. The complexity of the format could cause additional preservation risks. Complexity here means the compatibility between different format versions, semantic information necessary for correct rendering, using of compression, missing standard or documentation. In our list as complex formats were marked “PDF”, “HTML”, “TIF”, “SXW” and “DXF”. The dissemination level plays an important role in development of associated software tools and popularity of the format. In this regard high preservation risk have “MAC”, “SXW” and “DXF”. Some formats in the associated literature and in expert community are marked as outdated or deprecated due limited using of this format or its some of its versions. These formats are “DOC”, “HTML”, “PPT”, “MAC”, “SXW” and “DXF”. The open or standardised formats have lower preservation risks like “GIF”, “PNG”, “MP3”, “PDF”, “JPG” and “HTML”. Formats that have homepage have lower risks due to additional information placed in their homepages. Our tool found homepages for three formats “PDF”, “GIF” and “BMP”. These formats therefore are regarded as having lower risks. The genre information also reduce risks for “GIF”, “PNG”, “MP3”, “PDF”, “JPG”, “DOC”, “HTML” and “TIF”. The creation date factor could be implemented in different ways. In our meaning the older is the file format the more it was used and the more stable it is. Therefore “GIF”, “PNG”, “MP3”, “PDF”, “DOC”, “HTML” and “TIF” have low risk expectation in this regard. Other researcher could consider the latest date as more reliable. Important aspect for digital preservation is an ability to migrate file from one format to another. In this regard all of examined files have low risk in regular institutional environment. Digital rights information plays increasingly important role in digital preservation. Extraction of this important information is a topic of future work. Publisher and creator information gives us additional source to decide how much trust we will give to the given publisher. Our risk analysis tool found needed information for “MP3”, “DOC”, “HTML”, “PDF”, “GIF”, “BMP” and “JPG”. In order to evaluate how frequently particular format is used in libraries preservation workflows we make use of the expert knowledge. The most popular formats are “GIF”, “PNG”, “MP3”, “PDF”, “JPG”, “DOC”, “HTML”, “TIF”, “BMP” and “PPT”. In order to accumulate expert knowledge like in case of frequently used formats we are working on designing our own data repository that provides information missed in other LOD repositories. Similarly the compression support, web browser support and vendor support information is a topic of future work.

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” were retrieved genre, creation date, publisher and creator information, whereas these factors are missing for “PPT”. This case does not mean that such information does not exist. That only demonstrates that it is not included or not found in LOD repositories. The same consideration is valid for the “software count” value 12 of “MP3” format. As we know there should be much more associated software tools that are able to handle this format.

At that point we should state that we didn’t analyze all formats and that evaluated results currently require verification by human expert and further optimisation of calculation methods. Evaluation results presented in Figure 5 and Table 1 are limited to the information automatically collected from mentioned above LOD repositories and is customized by applied expert rules. Therefore these results can’t be regarded as absolutely accurate. The classification settings for risk factors are 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.

The figure 5 demonstrates the distribution of the analyzed file formats according to their evaluated risk scores. The most reliable formats are marked by the green color, the middle risk formats with orange color and the formats with the most risk are flagged by the red color. Each format is also marked by its risk score in percent. Concluding this part we evaluated that “GIF” (6), “PNG” (10), “MP3” (10),



| File Format Extension | Total Risk Score | Total Risk Level |
|---|---|---|
| pdf | 0.14 | Low |

Detailed List of Format Risk Scores

| Risk Factor | Risk Value | Weight | Risk Score | Weighted Risk Score | Risk Level |
|---|---|---|---|---|---|
| Software Count | 28 | 1.0 | 0.3 | 0.3 | Middle |
| Vendors Count | 2 | 1.0 | 0.0 | 0.0 | Low |
| Versions Count | 17 | 1.0 | 1.0 | 1.0 | High |
| Has Description | 2 | 1.0 | 0.3 | 0.3 | Middle |
| Has MIME Type | true | 0.2 | 0.0 | 0.0 | Low |
| Format Existence Period | true | 1.0 | 0.0 | 0.0 | Low |
| Format is Complex | true | 1.0 | 1.0 | 1.0 | High |
| Format is Wide Disseminated | true | 1.0 | 0.0 | 0.0 | Low |
| Format is Outdated or Deprecated | false | 1.0 | 0.0 | 0.0 | Low |
| Has Genre | true | 0.5 | 0.0 | 0.0 | Low |
| Has Homepage | true | 0.5 | 0.0 | 0.0 | Low |
| Format is Open (standardised) | true | 1.0 | 0.0 | 0.0 | Low |
| Has Creation Date Information | true | 1.0 | 0.0 | 0.0 | Low |
| Is File Migration Supported | true | 1.0 | 0.0 | 0.0 | Low |
| Has Digital Rights Information | false | 0.3 | 1.0 | 0.3 | High |
| Has Publisher Information | true | 0.1 | 0.0 | 0.0 | Low |
| Has Creator Information | true | 0.1 | 0.0 | 0.0 | Low |
| Frequently Used (popular) | true | 1.0 | 0.0 | 0.0 | Low |
| Is Compressed File Format | false | 0.9 | 0.0 | 0.0 | Low |
| Is Supported By Web Browsers | true | 0.5 | 0.0 | 0.0 | Low |
| Is Supported By Important Software Vendors | true | 0.3 | 0.0 | 0.0 | Low |

| File Format Extension | Total Risk Score | Total Risk Level |
|---|---|---|
| tif | 0.26 | Middle |

Detailed List of Format Risk Scores

| Risk Factor | Risk Value | Weight | Risk Score | Weighted Risk Score | Risk Level |
|---|---|---|---|---|---|
| Software Count | 135 | 1.0 | 0.0 | 0.0 | Low |
| Vendors Count | 1 | 1.0 | 0.3 | 0.3 | Middle |
| Versions Count | 9 | 1.0 | 1.0 | 1.0 | High |
| Has Description | 2 | 1.0 | 0.3 | 0.3 | Middle |
| Has MIME Type | true | 0.2 | 0.0 | 0.0 | Low |
| Format Existence Period | true | 1.0 | 0.0 | 0.0 | Low |
| Format is Complex | true | 1.0 | 1.0 | 1.0 | High |
| Format is Wide Disseminated | true | 1.0 | 0.0 | 0.0 | Low |
| Format is Outdated or Deprecated | false | 1.0 | 0.0 | 0.0 | Low |
| Has Genre | true | 0.5 | 0.0 | 0.0 | Low |
| Has Homepage | false | 0.5 | 1.0 | 0.5 | High |
| Format is Open (standardised) | false | 1.0 | 1.0 | 1.0 | High |
| Has Creation Date Information | true | 1.0 | 0.0 | 0.0 | Low |
| Is File Migration Supported | true | 1.0 | 0.0 | 0.0 | Low |
| Has Digital Rights Information | false | 0.3 | 1.0 | 0.3 | High |
| Has Publisher Information | false | 0.1 | 1.0 | 0.1 | High |
| Has Creator Information | false | 0.1 | 1.0 | 0.1 | High |
| Frequently Used (popular) | true | 1.0 | 0.0 | 0.0 | Low |
| Is Compressed File Format | true | 0.9 | 1.0 | 0.9 | High |
| Is Supported By Web Browsers | true | 0.5 | 0.0 | 0.0 | Low |
| Is Supported By Important Software Vendors | true | 0.3 | 0.0 | 0.0 | Low |

**Figure 6: Sample risk reports for PDF and TIF file formats.**

"PDF" (14), "JPG" (21), "DOC" (21) and "HTML" (24) have the lowest risk in percent. "TIF" (26), "BMP" (26) and "PPT" (34) have a middle preservation risk. The "MAC" (38), "SXW" (48) and "DXF" (49) formats was estimated as having the most risk. "BMP" and "TIF" have the same middle risk value scored by 26 percent but each of these formats has different weighted risk components. If we break down the resulting risk in risk factors we will see that "TIF" has more descriptions, but is more complex then "BMP". The genre information for "BMP" was not found, whereas "TIF" is missing homepage link. Creation date was detected only for "TIF". In contrast to "TIF" the "BMP" format has publisher and creator information. "TIF" supports compression what increases the preservation risk.

### 4.3 Web service for risk analysis report

In order to accumulate user feedback and to improve the approach the presented functionality was implemented in form of a web service (see Figure 6). In this picture we present a risk analysis report for "PDF" and "TIF" file format extensions. The "PDF" format has the low preservation risk with 14 percent and the "TIF" format has the middle preservation risk with 26 percent. The report comprises risk value, weight of the particular risk factor, calculated risk score, weighted risk score and risk levels in textual form. The most significant risk factors like software count, vendors count, versions count, standardisation, popularity, description factor, creation date factor and migration factor have the highest weight 1.0, the less important factors have weights in range between 0.1 and 0.5. The risk analysis reports provided by Web service demonstrate that our hypothesis was correct and file format data automatically aggregated from LOD repositories provides the rule engine with sufficient information and enables risk estimation for different file formats. The distribution of calculated format risk scores also proves that file formats flagged as "low risk" formats are most reliable file formats for digital preservation workflows. And old, outdated formats like "SXW" or "DXF" demonstrate high preservation risk.

## 5. CONCLUSIONS

Within this paper we presented the risk analysis service for file formats which employs FFMA knowledge base with rich descriptions of computer file formats. The service uses semi-automatic information extraction from the LOD repositories, analyzes and aggregates knowledge that facilitates decision making in different institutions for preservation planning. The main contribution of this paper is a definition of the risk factors with associated severity level and its automatic computation based auf information from knowledge base created from LOD repositories. The FFMA knowledge base is created using the ontology mapping approach for collecting data from LOD repositories. This allows automatic retrieval of rich, up to date information reducing so the setup and maintenance costs for the risk analysis service. The evaluation of preservation friendliness is based on risk scores computed with the help of expert models. Risk metrics are customizable by configuration files and provide risk estimation parameter that are exactly adjusted for particular user. We integrated FFMA tool with expert rule engine for automatical format risk analysis and decision support. Open repositories (PRONOM, DBPedia and Freebase) ontologies mapping concerning digital preservation domain facilitates mapping between formats, software and vendors and supports efficient search for giving input format. Customized File Format Ontology and rules ensure information extraction quality assurance. An important contribution of this paper is the exploitation of the up to date open knowledge bases for aggregating information within the recommender's knowledge base, reducing so the setup and maintenance costs for the FFMA. Since the knowledge acquisition and aggregation process is fully automated, this will allow to easily upgrade the knowledge base and the risk rules engine. The scalability of information extraction was improved by reducing of domain knowledge acquisition efforts by means of aggregated information storing in a database. Web service[9] was implemented to support knowledge base manage-

[9] http://ffma.ait.ac.at:8080/preservation-riskmanagement/

ment and decision making based on risk analysis report for file formats. We employed AI technologies (i.e. expert rules) for inferring explicit knowledge on the nature and preservation friendliness of the file formats. A statistical analysis of the aggregated information and the qualitative analysis of the aggregated knowledge are presented in the evaluation part of the paper. As future work we plan the extension of expert rules for format risk analysis with new aspects in order to increase risk estimation accuracy. We also plan using of additional knowledge sources (e.g. vendor's web sites, further knowledge bases) and request fields for extending the knowledge related to the software tools, vendors and their relationship to the existing file formats.

## 6. ACKNOWLEDGMENTS

This work was supported in part by the EU FP7 Project SCAPE (GA#270137) www.scape-project.eu and partially by the EU project "ASSETS - Advanced Search Services and Enhanced Technological Solutions for the European Digital Library" (CIP-ICT PSP-2009-3, Grant Agreement n. 250527). The authors wish to thank Paul Wheatley from the British Library for his thoughts on the topic.

## 7. REFERENCES

[1] B. Aitken, P. Helwig, A. Jackson, A. Lindley, E. Nicchiarelli, and S. Ross. The planets testbed: Science for digital preservation. *Code4Lib*, 1(3), 2008.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[3] 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.

[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 Workshop on Human Decision Making in Recommender Systems (Decisions@RecSys'11) and User-Centric Evaluation of Recommender Systems and Their Interfaces-2 (UCERSTI 2) affiliated with the 5th ACM Conference on Recommender Systems*, 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] 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.

[7] 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.

[8] L. Jens, S. Jörg, and A. Sören. Discovering unknown connections -the dbpedia relationship finder. In *Proceedings of the 1st Conference on Social Semantic Web (CSSW)*, volume P-113, pages 99–109, Leipzig, Germany, 2007. Gesellschaft für Informatik.

[9] R. King, R. Schmidt, A. Jackson, C. Wilson, and F. Steeg. The planets interoperability framework: An infrastructure for digital preservation actions. In *ECDL09 Proceedings of the 13th European conference on Research and advanced technology for digital libraries*, volume 5714/2009, pages 425–428. Springer-Verlag, 2009.

[10] B. Kurt, E. Colin, P. Praveen, S. Tim, and T. Jamie. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD '08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1249, New York, NY, USA, 2008. ACM.

[11] 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.

[12] D. S. Rosenthal. Format obsolescence: assessing the threat and the defenses. *Library Hi Tech*, 28(2):195–210, 2010.

[13] J. Rothenberg. Digital preservation in perspective: How far have we come, and what's next? *Future Perfect 2012*, 2012.

[14] 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.