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 the automatic estimation of preservation risks for file formats. The main contribution of this work is the definition of risk factors with associated severity levels and their automatic computation. Our goal is to make use of a solid knowledge base automatically aggregated from linked open data repositories as the basis for a risk analysis in the digital preservation domain. This method is meant to facilitate decision making with regard to preservation of digital content in libraries and archives. We have developed a tool for aggregating rich and trusted file format descriptions. It exploits available linked data resources and uses expert models to infer knowledge regarding the long-term preservation of digital content. 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 techniques (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

Preservation planning activities make use of the analysis and evaluation of file formats used for encoding digital content. The preservation risks for a particular file format are difficult to estimate and the definition of risk factors is still an open research topic. Intensive human expert involvement is required for searching and aggregating information about preservation risks and estimating of their possible impact in the future [2, 18]. The definition of risk factors for long term preservation can vary depending on preservation goals, workflows and assets used by a particular organisation. Also, the classification and weighting of risk factors is a challenging task, and is strongly dependent on the level of knowledge and experience of human experts. Individual domain specific knowledge bases do not contain all necessary semantic information required to perform an estimation of the preservation risks. The richness and the quality of knowledge base plays an important role in taking decisions on preservation planning. Even though the world wide web has turned out to be the largest knowledge base, the published information lacks a unified well-formed representation. The linked open data (LOD)[1] and Open Knowledge[2] initiatives address these weaknesses by defining guidelines for publishing structured data in standardized and queryable format. In order to aggregate sufficient knowledge about file formats for risk analysis we link 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 formats and therefore plays a fundamental role in the process of managing electronic records. Many file formats are properly documented, are open-source and well supported by producer. Other formats may be outdated, changed 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) ([7]) system depicted in Figure 1, which aims at preparing the ground for knowledge base recommenders like DiPRec [6]. FFMA reuses the experience of building preservation planning tools and addresses the topic of digital long-term preservation. It performs an analysis of file formats based on the concept of risk scores. The knowledge base is built by following a linked data approach. Concretely, the information regarding file formats, software tools and vendors is 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 related to the usage of ontology mapping (see Figure 1) for the integration of different linked data sources into a common knowledge base. Decision support based on the 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 information on file formats using a (semi-) au-

[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/

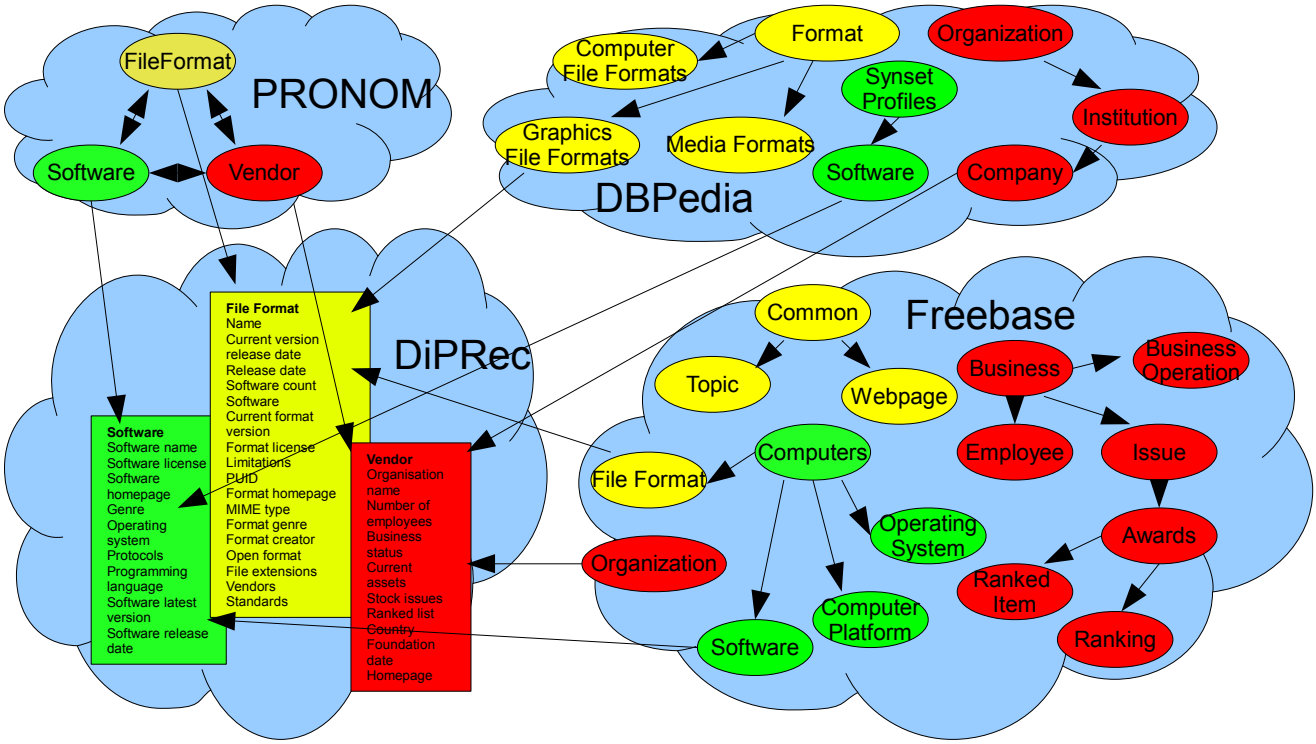

Figure 1: PRONOM, DBPedia and Freebase digital preservation domain related ontology sections mapped to the DiPRec file format ontology.

tomatic approach for knowledge extraction from the linked data repositories independent from the query language supported by individual repositories. We aim at designing well structured knowledge base with defined rules and scored metrics that is intended to provide decision making support for preservation experts. The paper is structured as follows: Section 2 gives an overview of 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 an outlook about planned future work.

2. RELATED WORK

In [10] Andrew Jackson evaluated competing hypotheses regarding software obsolescence issue employing format identification tools for selecting appropriate preservation strategies. One of these hypothesis is presented by Rothenberg [17] and emphasizes that all formats should be considered brittle and transient, and that frequent preservation actions will be required in order to keep data usable. In contrast to that hypothesis the Rosenthal [16] 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. 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, etc. In [10], the author 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 PANIC tool [9] had the goal to 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 was to aggregate data and metadata for further analysis, but this information is not easy browseable and the size of the repository is relative small in comparison to the LOD sources. Also there is no common understanding in the community about the meaning of term “obsolete” as mentioned above.

The AONS II tool [15] aimed at identifying file formats used for encoding digital collections, retrieving information regarding obsolescence risk indicators. The tool was building collection profiles and was referencing external format registries. This tool was able to distinguish accurately between

different versions of formats, in order to identify relevant risk levels. AONS II tool struggled 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 tried to map it to the likely matching format identifiers used by external registries. But this tool did not apply risk factor metrics for risk calculation. Inspired by [15] we realized the need to develop a central web service that shares the results of local risk assessments with the community of interest. We aim at defining risk metrics based on experience of community members which share their individual expertise on defining and identifying risk factors. This would allow LOD registries to leverage the experiences and expertise of the contributing preservation community and add considerably to their usefulness.

The goal of the SPOT (Simple Property-Oriented Threat) model [18] is to 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 for preserving digital content, and threats for 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 exploit open knowledge from LOD repositories.

In the proposed approach we do not intend to mark down obsoleted formats, since there are different hypotheses and no common accepted definition for format obsolescence. Therefore we do not intend to treat obsolescence in a generalized form, but we treat it in an contextualized one. We define 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. With FFMA 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", "is migration supported", "has digital rights information", etc. Most of these factors have influence on rendering the content. FFMA has the advantage of enabling users to configure the risk factors and scores according to their institutional context.

The format risk analysis approach in [5] presents the P2 registry, which is an RDF-based framework. The P2 registry employs information containing in DBPedia and PRONOM repositories and supports its own format risk analysis system. The main goal of the P2 platform is 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 makes use of the rich Freebase repository as well and provides a modular architecture capable to easy integrate further repositories, even if they are not RDF based.

Aditionally, FFMA uses a rule engine for risk analysis that handles further risk factors not covered by P2 registry and also supports their customization. Additional expert rules can be simply added to the model concept and the weighting severity levels are customizable as well.

Existing tools for long term preservation planning like Plato ([12, 1]) enable different digital preservation actions like identification, characterization and content 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. The Plato tool defines decision criteria [3] for formats depending on institutional risk profile, but these criteria mainly are concentrated on format properties that can be obtained from P2 fact base and have predefined property values in contrast to normalized numerical values in FFMA expert system.

3. THE RISK ANALYSIS PROCESS

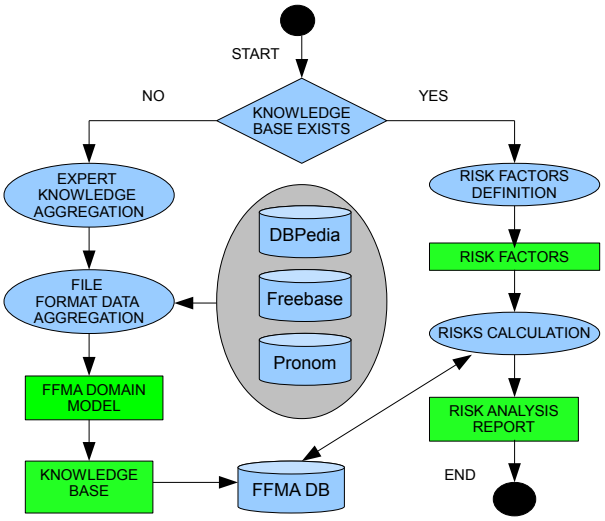

Figure 2: The format risk analysis workflow.

Figure 2 presents the risk analysis workflow. The building of the knowledge base (i.e. left side of the sketch) is a prerequisite for performing the risk computations. This includes the acquisition of expert knowledge and an aggregation of rich file format data. The creation of risk analysis reports is a two-step process based on the definition of risk factors and the computation and interpretation of risk scores (i.e. right side of the sketch). The result of risk calculation is presented in HTML format. The extended description of individual steps within this process is presented in the following sections.

3.1 Aggregation of File Format Data

The FFMA module for aggregation of file format descriptions collects information from LOD repositories and enhances it by using the expert knowledge aggregation module. At runtime, the aggregated metadata is processed and represented according to the underlying FFMA domain model

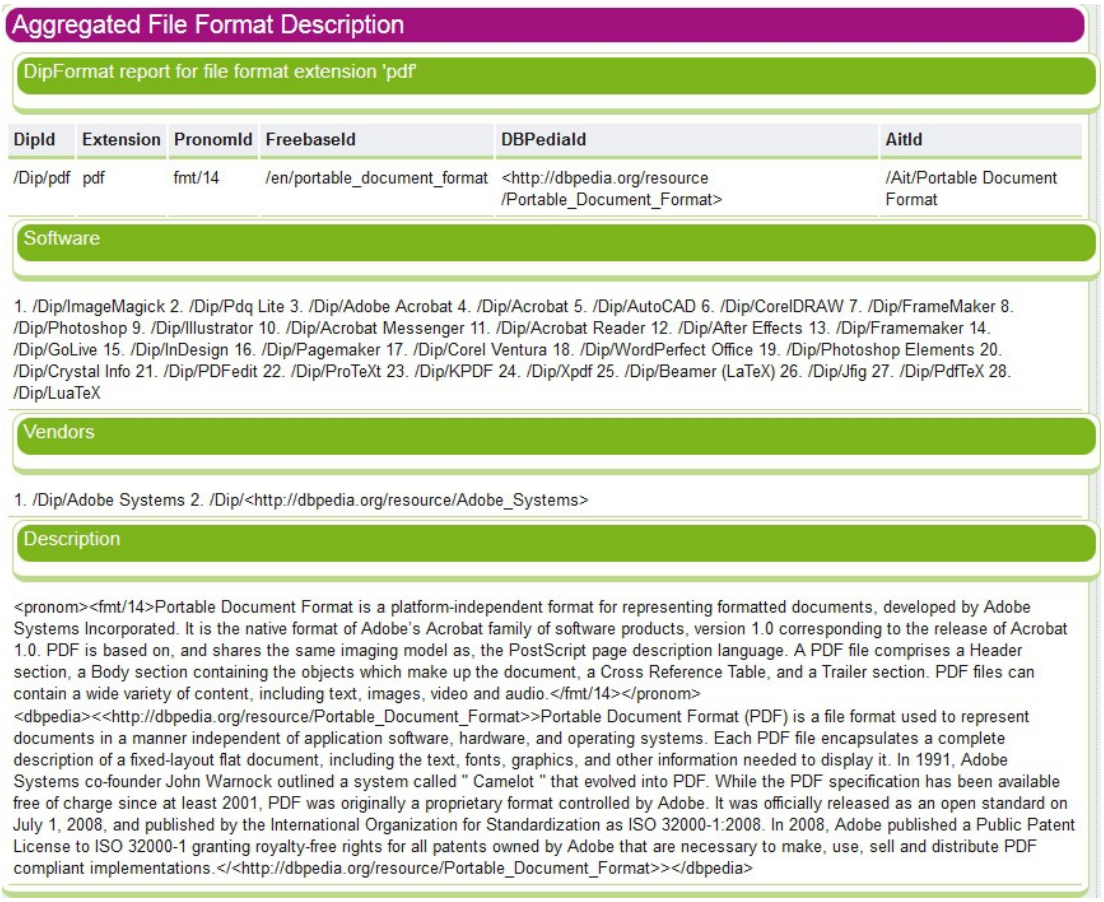

Figure 3: Example of an aggregated data report for PDF file format.

by taking in account the configurations for a specific exploitation context. These configurations define 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 after ontologies mapping in Expert Knowledge Aggregation module. The acquired domain knowledge in stored in a local database and further used in the reasoning risk computation process. We consider Freebase [13] as one of the most valuable sources for information extraction. It is a practical, scalable semantic database for structured knowledge. The PRONOM data format looks very similar to the FFMA ontology classes but it doesn't contain all necessary properties (like genre or vendor business status) that DiPRec requires to incorporate significant data from another ontologies. Extending the PRONOM repository, we collect information from additional sources and aggregate it in a homogeneous representation in the FFMA knowledge base, by using the FFMA domain model. The assignment to given property sets, the functions for value normalization, the queries for specific LOD repositories are the main constituent parts of the property definition model. An example of aggregated description for PDF format is presented in Figure 3. 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. DBPedia allows to post sophisticated queries using SPARQL query and OWL ontology languages [11] 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) [4]. PRONOM data is released as LOD and is accessible through a public SPARQL endpoint.

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 specific database areas. This model increases performance because we do not need to perform expensive database queries with every operation. 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. File format properties are designed to give an option at hand for

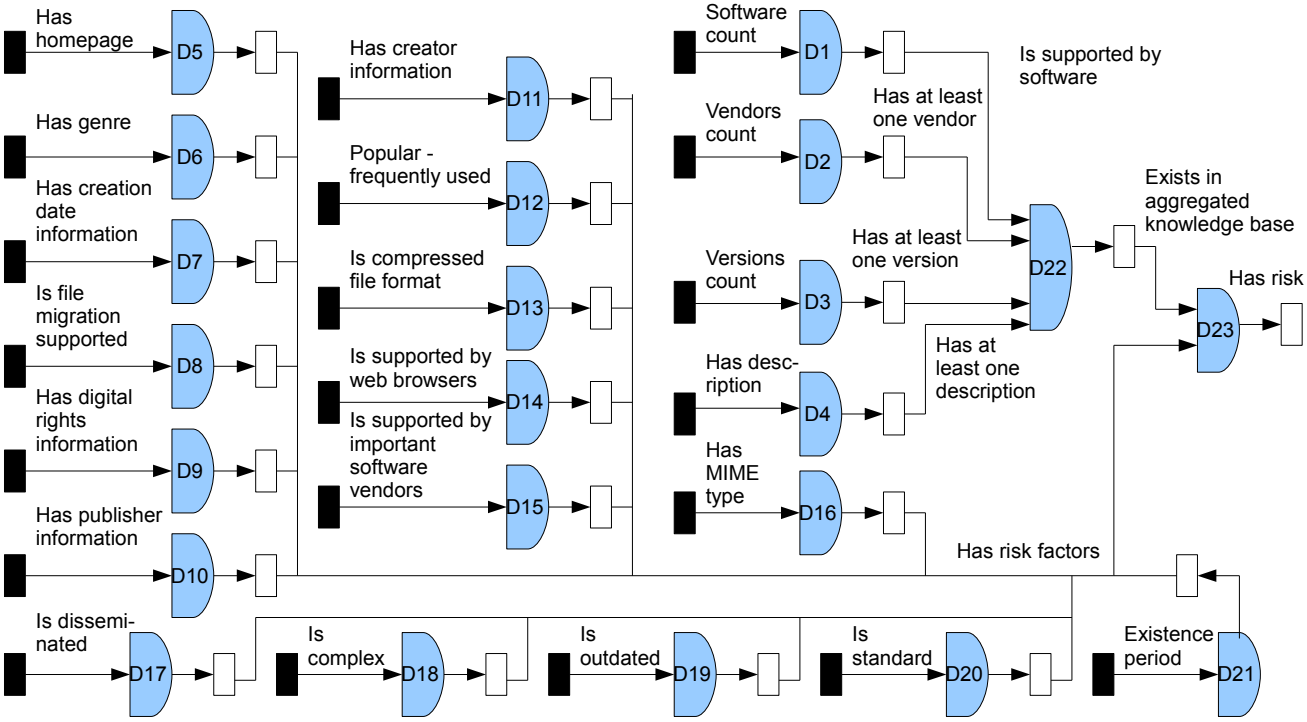

Figure 4: Forward rule chaining for risk analysis.

definition of user rules, metrics and classifications. The risk factors are used to compute overall preservation risks for a given file format.

3.2 Risk Factors Definition

The evaluation of the completeness of the knowledge aggregated from data registries (e.g. the percentage of file formats for which the genre property is available) gives some rough estimates about which risk properties should/can be defined and how to interpret them (i.e. weight and severity assignments). The most significant data repository queries in terms of digital preservation addresses PRONOM Id (i.e. PUID), file formats, software and software vendors. The properties of these main classes including computer platform, genre, license, programming language, release date, homepage, compression type and so on are of interest for risk analysis. Optionally, the user is able to extend the default risk analysis model by defining its own property sets of inferred knowledge and classifications using correspondent configuration files.

The information obtained from the digital preservation domain experts and from conducted experiments must be well structured. Typical scenarios were defined and the parameters used by library experts for collection handling were identified. Then linguistic labels were defined to classify measured values of each parameter and associated ranges. Finally, were determined 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 [14] [8].

The most significant risk factors are related to the availability of software tools and vendors providing support for a particular file format (see Figure 4). For example, the version count metric could be interpreted in different ways. On the one hand the more versions a format has the more work is invested in its development and support. This implies that the given format is in use and well supported. On the other hand with the version count increases the probability that different versions will increase complexity and might generate conflicts when designing digital preservation workflows (e.g. for format migration purposes).

By changing severity values and classification settings, each customer could adjust the meaning of this risk factor for his specific context, 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. The digital rights play increasingly important role in digital preservation. 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 between different versions of the same format, compression factor etc. Because the expert system contains information not only about format extensions but also about different versions, this knowledge could be covered by separate rule. Some formats are implicitly or explicitly declared as outdated or deprecated. The standardized formats have better chances of having a long time support. The time passed from the first release of a format is an additional metric for

File Format	Overall Risk Score	Overall Risk Level
pdf	0.14	Low

Detailed List of Format Risk Scores

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

File Format	Overall Risk Score	Overall Risk Level
tif	0.26	Middle

Detailed List of Format Risk Scores

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

Figure 5: Sample risk reports for PDF and TIF file formats.

risk estimation. Mature and popular formats present lower preservation risk. Software, vendors and versions count factors together with a description factor build an aggregated rule whether the given format is supported by FFMA. Missing one of these important pieces of information means that the regarded LOD repositories do 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. The forward rule chaining for file format analysis is presented in Figure 4. Forward chaining is the process of moving from the antecedents ("if" conditions) to the consequents ("then" actions) in a rule-based system. A specific rule is triggered if all of its inputs are available (i.e. a risk is present only if all assigned input properties are available). The antecedent is considered satisfied when the input values match the assertion, in which case the rule computes a risk value as consequent, otherwise a default risk value is set as consequent. Assertions are depicted by black rectangles on the input side and by the white rectangles on the output side (i.e. as result of the rule evaluation). The rules are presented by blue half-spheres, respectively. The output of one rule is used as an input for the following rule in the chain.

As an example, the rule-base system may start risk identification with the rule D1 supposing that software count is higher than 0. If the antecedent pattern defined in classification settings matches that assertion, the value x becomes "is supported by software" and the rule D1 fires. When 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 similar iteration through remaining rules.

3.3 Risk Computation

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. The computation and interpretation of risk scores is completed within the Risk Calculation task (see Figure 2) by using the previously presented forward chaining model (see Figure 4). The risk score for a particular property is evaluated from risk analysis model dependent on metrics, property weight and risk interpretations. Each rule is responsible for the computation of a risk factor, and the weighted risk scores are used for computing the total risk score for a given format (see Figure 5 for an illustrative example).

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 [6] and is computed as a weighted sum over all risk factors:

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

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 $d(p, PFV(p))$. The 'preservation con-

form' values and the metrics for distance computation are specified within the property definitions.

The final risk report contains detailed information about computed risk scores for each property, the weighting factors used in risk computations, the total risk scores for a file format and their user friendly interpretations (i.e. indication of severity levels). This kind of report provides a solid evaluation of the file format descriptions and estimates the preservation friendliness based on the interpretation of computed preservation risks.

4. EXPERIMENTAL EVALUATION

The evaluation of format risks was conducted with the FFMA knowledge base aggregated for development of DiPRec recommender. 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. It is expected that the distribution of calculated format risk scores will match to the associated information that was found 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. 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 basis of this service relies on rich data descriptions retrieved from LOD repositories. 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 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. Service provides generation of rich format descriptions and a report on format risks.

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.

Aggregated data reports are presented in HTML format by the FFMA service. An example for the PDF file format is presented in Figure 3. This report comprises the FFMA identifier */Dip/pdf*, the unique identifiers within external repositories describing the *pdf* format. According to the LOD principles, each linked data repository has its own mechanism for non-ambiguous referentiation of the managed entities represented by unique Web URLs. By having a reference in a correct format, a user is able to easily request the information from a web service. In this case, the PRONOM identifier is *fmt/14*[6], the Freebase one is */en/portable_document_format*[7] and DBPedia is *Portable_Document_Format*[8], respectively.

Additionally information about 28 different software tools and one vendor associated with this file format was aggregated and presented by their 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 real-time performance (lower then a half of second on regular PCs). Aggregated reports on file formats contain information like "FileFormatDescription", "SoftwareName", "RepositoryName", "SoftwareHomepage", "SoftwareDescription" etc. FFMA returns evaluated software, vendor and risk report objects in HTML format. The processing of LOD objects supports storage, retrieval and analysis of information retrieved from Web repositories. This structured information is a knowledge base to be used for deriving preservation recommendations.

4.2 Computation of risk factors

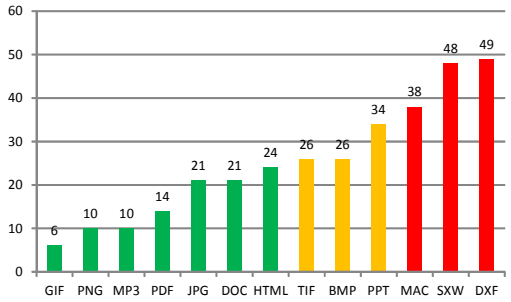

Figure 6: The distribution of the file formats with associated risk scores in range from 0 to 100 percent

Figure 6 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 yellow color and the formats with the highest risks are flagged by the red color. Each format is also marked by its risk score in percent. In consequence, the experimental evaluation shows that *GIF* (6), *PNG* (10), *MP3* (10), *PDF* (14), *JPG* (21), *DOC* (21) and *HTML* (24) present the lowest preservation risks. The *TIF* (26), *BMP* (26) and PPT (34) formats have a middle preservation risk, while the *MAC* (38), *SXW* (48) and *DXF* (49)

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

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

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

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

formats were evaluated as being the less trusted ones. *BMP* and *TIF* have the same overall risk score by 26 percent, but this is a result of aggregating the weighted scores of different low-level risks. By breaking down the results in risk factors, one can verify that the *TIF* format has more descriptions, but in the same time it is more complex then the *BMP*. The genre information for *BMP* was not found (i.e. in the aggregated knowledge base), whereas for *TIF* no homepage link is available and a creation date was available only for *TIF*. In contrast to this, for *BMP* format the publisher and creator information is available. Additionally the *TIF* format is a compressed one, fact that increases its preservation risk.

The aggregated risk scores were computed by using the model described in Section 3 by employing the information aggregated within the knowledge base and computing individual risk factors relevant to the given file formats. Table 1 presents an overview of the computed low level risks for the formats included in the evaluation set. 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 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.

Therefore, the risk regarding the "software count" for *SXW* was considered as being high, the risks for *DOC*, *HTML*, *TIF* and *MAC* extensions as low and medium risk is associated with remaining formats. By defining classifications for this risk factor, it was expected 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, the user may redefine classification settings according to his risk estimation preferences. The lowest risk for "vendors count" risk factor were calculated for *GIF*, *MP3* and *PDF* formats with two to three vendors. The remaining formats have middle value associated with this risk, in consequence no high risk regarding "vendors count" component was detected for the given data set. High vendor risk would be expected in the case that no vendors were documented for particular format. It was assumed 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). Regarding availability of textual descriptions, it was expected that the more information was found, the lower is the risk. According to this risk definition the high risks were detected 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 there is no low risk among them. The MIME type is an essential reference in order to address a file format and to create a connection between different file format ontologies or identification tools. Most of the presented formats have found an associated reference. Only three formats are missing the MIME type: the *MAC*, *SXW* and *DXF* formats. The longevity of the format existence period could give us 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 to limited using of this format or 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 researchers could consider the latest date as more reliable. Another 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 should be given to the particular publisher. Our risk analysis tool found the information required for *MP3*, *DOC*, *HTML*, *PDF*, *GIF*, *BMP* and *JPG*. In order to evaluate how frequently particular format is used in libraries preservation workflows was used 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 was designed new 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* 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 Figure 6 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 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 Web service for risk analysis report

Finally, the presented approach was implemented as a REST-Full web service, allowing individuals and third party applications to make use of available risk computations[9]. We aim also at collecting more user feedback and to improve the presented risk computation models. Figure 5 presents user friendly presentations of the analysis reports regarding the *PDF* and *TIF* file formats. The *PDF* format has the low preservation risk with 14% and the *TIF* format has the middle preservation risk with 26%. The report includes the nominal values for the risk properties, their weighting in risk computations, the derived risks scores, the individual interpretations (i.e. risk level) and their weighting for the computation of the total risk score. In the provided examples, 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. The file format descriptions automatically aggregated from LOD repositories provide sufficient information to enable estimation of preservation risks for various file formats. The distribution of calculated format risk scores proves that file formats flagged as "low risk" formats are (still) reliable file formats. Old, outdated formats like *SXW* or *DXF* were identified as presenting increased preservation risks by the given models.

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

5. CONCLUSIONS

This paper presents the risk analysis service which employs FFMA knowledge base with rich descriptions of computer file formats. The service uses semi-automatic information extraction from the Linked Open Data repositories, analyzes and aggregates knowledge that facilitates decision making in different institutions for preservation planning. The main contribution of this paper is the definition of the risk factors, their automatic computation and interpretation based on aggregated knowledge base. 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 the setup and maintenance costs for the risk analysis service. Since the knowledge acquisition and aggregation process is automated, this will allow the aggregated knowledge base to be easily updated. The scalability of information extraction was improved by reducing the domain knowledge acquisition efforts by means of storing the aggregated knowledge in a local database. The evaluation of the preservation friendliness is based on the expert models employed for performing the computation of risk scores. The underlying expert model can be easily adapted to the preservation requirements of particular institutional contexts through the customization of the configuration files, the risk definitions and their associated severity levels. A Web service was implemented to support the evaluation of the aggregated knowledge base and to support decision making on digital preservation actions based on the provided risk analysis reports. The evaluation part of the paper presets the computation of risk analysis reports for a representative set of 13 well known file formats. The presented model makes use of 23 different risk factors. The interpretation of experimental results demonstrates the viability of the proposed approach. Anyway, there are still two main drawbacks of the proposed approach. The first of them is related to the need to reason based on incomplete information (e.g. the description of file formats is not complete in either of the given repositories). The second one is related to the need to adjust the weighting of the risk factors according to individual institutional contexts.

As future work we plan using of additional knowledge sources (e.g. vendor's web sites, further knowledge bases) and additional properties for format descriptions (e.g. popularity of file formats available on `http://www.fileinfo.com/`). The extension of expert rules with new risk factors, improving the accuracy of the expert model and enhanced identification of software tools supporting individual file formats are additional research topics to be investigated.

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] 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.

[3] C. Becker and A. Rauber. Decision criteria in digital preservation: What to measure and how. *Journal of the American Society for Information Science and Technology (JASIST)*, 62(4):1009–1028, 2011.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] S. Hoorens, J. Rothenberg, C. van Oranje, M. van der Mandele, and R. Levitt. Addressing the uncertain future of preserving the past: Towards a robust strategy for digital archiving and preservation. Technical report, RAND Corporation, 2007.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] A. McHugh, S. Ross, P. Innocenti, R. Ruusalepp, and H. Hofman. Bringing self-assessment home: Repository profiling and key lines of enquiry within drambora. *International Journal of Digital Curation*, 3(2), 2008.

[15] 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.

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

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

[18] 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.