On Enhancing the FFMA Knowledge Base

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

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

ABSTRACT
Ensuring the long term access to digitized content is a major concern of digital libraries. The document migration and summarization are key activities employed reach this goal. The evaluation of preservation friendliness and making recommendations for long term preservation requires deep domain knowledge which is currently not available in any integrated knowledge base. In this paper we present an approach for enhancing the automatic aggregated knowledge on computer file formats. A clustering algorithm is employed to identify related file formats and to predict missing semantic associations between file formats and software tools. This is used to improve the discovery of software tools supporting the less popular file formats.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: System issues; H.3.3 [**Information Search and Retrieval**]: Clustering

Keywords
digital preservation, file format categorization, related file formats

1. INTRODUCTION
One important aspect of preservation planning is related to the file formats used for encoding the digital information. Currently, the information about the file formats is only partially available in domain specific knowledge bases and it is not appropriate formatted, accurate or complete in the open data repositories. The activities related to the preservation of digital content in libraries and archives are associated with high financial efforts, therefore the decisions about preservation planning must be taken by using rich, trusted, as complete as possible domain knowledge. There were significant efforts made in the last years within this research direction, but the systems built by now fail to effectively support preservation planning activities, mainly because they lack of a solid knowledge base (i.e. containing rich descriptions and contextual metadata related to available file formats)[9]. Typical preservation plans include migration of the content available in old file formats into formats that are preservation-friendly (e.g. well supported by standard hardware and software systems, appropriate for publishing on the web or on paper). One of the big challenges of preservation planning is to find the appropriate software tools that are available for executing the preservation plans, given the multitudediversity of available file formats, software tools and version incompatibilities. The migration pathways provided by PRONOM is limited, due to the fact that this information is manually collected by a relative small community. In contrast to this, the semantic web resources (i.e. DBpedia, Freebase) are supported by large communities, but they typically don't have a preservation related background. In consequence, these repositories contain rich descriptions of file formats, software tools and their vendors, but there is an extremely low coverage of the software to file format linking. This paper is a continuation of the work presented in [2] and it is intended to provide a solid knowledge base for the risk analysis module of the File Format Metadata Aggregator (FFMA) service [3]. The main contributions of this paper consist in employing clustering algorithms for identifying related file formats, making use of genre classifications and predicting missing semantic links between software tools and file formats.

The rest of the paper is structured as follows: Section 2 gives an overview on related work and concepts and Section 3 presents the domain specific issues related to the recommendation of digital preservation actions. Its subsections present the enhancements added to the FFMA knowledge base and the algorithms used within the proposed approach. Section 4 presents the setup, evaluation and the interpretation of the experimental results. Section 5 concludes the paper and gives outlook of the future work.

2. RELATED WORK
Preservation planning is one of the important topics in the digital preservation, which is one of the newest research fields of computer science. Within this context, tools like PLATO [4] were developed with the goal of creating preservation plans by scheduling different actions like identification, characterization and migration. It uses a cost based model for evaluating the effectiveness of document migrations and uses a knowledge base for storing facts about file formats and migration paths. Registries like P2 [9] and its successor LDS3 [8] concentrate on building a knowledge base using the linked data approach and computing the preservation risks for individual file formats. Similar to our approach these systems integrate information collected from PRONOM and DBpedia, but they do not compute any enrichments, classifications and do not predict missing semantic links. The unified digital format registry project (UDFR) developed a platform based on semantic web technologies, which allows editing descriptions for file formats that were imported from PRONOM and MIME media types repositories [1]. In extensions to simple metadata aggregation, the

approach presented in this paper uses artificial intelligence technologies for enrichment and reasoning on the formats descriptions. For inferring explicit knowledge on related file formats we employed the well known algorithm: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [6]. Furthermore text based information retrieval models are use for computing similarities between natural language descriptions [7].

3. APPLICATION DOMAIN AND RELATED ISSUES

The knowledge based recommendation technology is the natural solution chosen for implementing tools supporting the preservation planning activities. They typically make use of expert rules and have the goal to analyze the compatibility of the content repositories with the state of the art and future technical infrastructures. The risks of not being able to archive, render or publish digital objects with modern tools are estimated. In the following we present a simplified representation of the digital preservation recommendation problem by illustrating the core of the recommendation algorithms:

```
IF
    NotPreservationFriendly(in: Format A)
THAN
    FindPreservationFriendly(in: Format A, out: Format B)
    FindMigrationSoftware(in: Format A, in: Format B, out: Software S)
RECOMMEND
    MigrateContent(in: Format A, in: Format B, in: Software S,
                   in: Configuration C, in: File NPF, out File PF)
```

where the type of the input and output variables belong to: *Format* - file format used for encoding content, *File* - a digital file storing multimedia content, *Software* - software tool used for processing a given file and, the *Configuration* used by the software tools in data migration processes. Within the pseudo code displayed above one can identify the key research questions that need to be solved by digital preservation recommender system:

Computation of preservation friendliness. The preservation friendliness of a given file format can be estimated by analyzing its complete description. This depends on the type of the content (i.e. text, image, audio, video), institutional context (e.g. archiving vs. web publishing), being open or standardized format, being supported by major vendors, rendering and processing with open source software, etc. Advances on this research topic are presented in [3].

Identification of migration Software. Whenever the digital content is packaged in an obsoleted or inappropriate file format, it is recommended to migrate it to a new representation (i.e. encoding) that is compatible with the modern communication technologies and processing/rendering software. As this information is not explicitly available, either in (open) domain specific knowledge base nor in semantic web. We aim at discovering this important information by using two heuristics: a) A software that is able to process two different file formats is able to convert between the two encodings (e.g. typically accessible through "Save as.." action) b) Each software is meant to process a group or related or equivalent file formats (i.e. document processors, graphic software, multimedia software, etc.). Our efforts for automatic clustering of the related file formats are presented in Section 3.2.

Preparing migration configuration. The conversion of the content from one encoding into another one requires provision of encoding specific and software specific parameters. This is achieved by evaluating the software tools and experimenting with them for ensuring the required quality of the conversion. This research topic is addressed by the work carried out in projects like Planets [4] and SCAPE [5]. In the current paper we focus our attention on the second research issues and we aim at identifying candidate software tools that are able to open specific file formats. The proposed approach uses the genre classifications and the free text descriptions to discover similarities between file formats and to infer predictions on matching software products. The currently used algorithms are not able to provide a high level of confidence, since the software and the file format versions are not taken in account. This is due to the fact that the version information is not available in linked open data repositories, except for a very few items.

3.1 Enhancing the FFMA Knowledge Base

A detailed analysis of the content and the size of the FFMA knowledge base was presented in [2]. It contains rich descriptions of about 594 file formats, 3719 software tools and 63 vendors aggregated from PRONOM, Freebase and DBpedia repositories. Despite of the richness of individual item descriptions, one of the weaknesses of the FFMA knowledge base is the low coverage of file format to software tools linking as presented in Figure 1. This histogram shows the distribution and the coverage of the file format to software relationships in the aggregated database. There are 154 file formats for which no software is known and there are 474 software tools for which no more that 3 supported file formats are known. From digital preservation point of view,

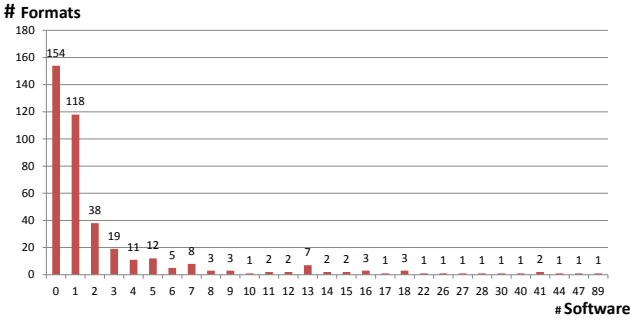

Figure 1: Histogram of software tool support for file formats

it is relevant how well a file format is supported, on how many platforms and how many software tools may render or edit it. In practice, the software tools are able to open more (related) file formats with different version (e.g. the most popular image file formats can be opened by the most of the image processing tools). By using a *linking through clustering approach* we aim at discovering important knowledge used within the preservation planning activities. For many software tools, there only a part of the list with the compatible file formats linked in the knowledge base, but there is a good chance that the tool is able to process additional similar file formats. For example, knowing that an graphic tool is able to process JPEG2000 files, there is a great chance that this tool will be able to process related file formats, like regular JPEG, Bitmap or TIFF. By using this reasoning, we aim at enhancing the digital preservation recommenders and enabling diagnosis in case that no migration solutions are provided. In this case, a set of candidate

software is generated including tools supporting related file formats (e.g. having similar genre and similar textual descriptions). External resources (e.g. homepages) might be manually checked to identify if one of the candidate tools is able to perform the conversion and to improve the recommender's knowledge base.

3.2 Related File Formats

For computing the related file formats clusters we use a variant of the most representative clustering algorithms, namely DBSCAN [6]. The ideas behind this algorithm is that the points within the cluster are mutually density-connected, which means in our case, that Format A and Format B belong to the same cluster in the case that each of the formats indicates the other one as being a neighbour. The definition of the algorithm is generalized and it is abstracted from the computation of neighbourhoods (i.e. distance between points in vector space).

The proposed algorithm uses textual information to compute distances between file format descriptions [7]:

$$dist(Q,T) = 1 - sim(Q,T) = 1 - \sum_{t \in Q,T} tf \cdot ln\frac{N}{df}, \quad (1)$$

Where $dist(Q,T)$ represents the distance between query format description Q and the target format description T, which is the inverse function of the similarity between the formats $sim(Q,T)$. t stands for the terms found in both descriptions, N for the total number of format descriptions, while tf represents the term frequency within the target format description and the df represents the document frequency, respectively (i.e. in how many format descriptions the term t occurs). In the experimental evaluation we make use of the implementation provided through the "MoreLikeThis" handler available in Solr [1].

4. EVALUATION

The experimental evaluation was carried out by using the FFMA knowledge base and the genre classification of file formats available in Wikipedia. The goal of this evaluation was to show that the textual descriptions aggregated from linked data can be used to identify similar file formats. Furthermore, we evaluate the tool support on cluster level which provides input for enhancing the migration pathway generation.

4.1 Identification of related file formats

The identification of the related file formats is performed by using the algorithm described in the previous section. The results of the clustering is depicted in Figure 2 showing distribution of the file formats over the 29 clusters identified by our algorithm containing at least 5 members. The centroids of individual clusters are represented on the X axis, while the Y axis represents the amount of formats that belong to the given cluster. The largest clusters are represented by the following centroids: *doc*, *ace* and *dwg* with a member count of 70, 35 and 34 respectively. The *doc* labeled cluster contains the textual documents, *ace* stands for the archiving formats cluster and *dwg* (DraWinG) for standard raster formats. Clusters calculation evaluated 51 clusters with the

[1] `http://wiki.apache.org/solr/MoreLikeThis`

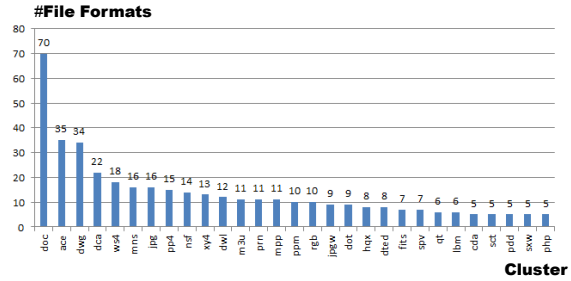

Figure 2: Distribution of file formats in clusters

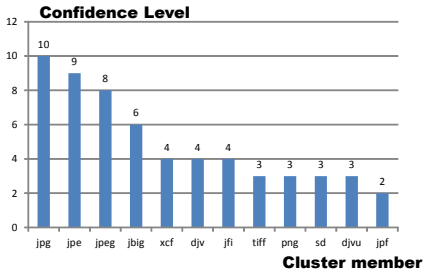

Figure 3: File formats in JPG cluster

nodes count in the range from 5 for *php* and 70 for *doc* cluster. Each cluster must have at least 5 members, otherwise the formats were considered as being outliers. Figure 3 presents the members of the *jpg*, most of them being well known raster graphics formats. The gain of the clustering

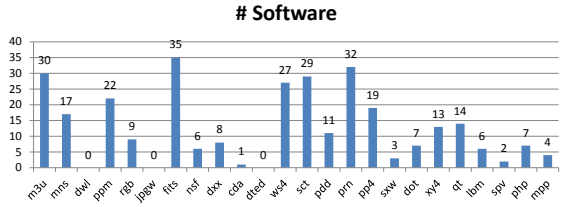

Figure 4: Software support for clusters.

consists in the identification of the software tools that are supporting several of the formats within the cluster. Figure 4 presents the association of the software support for the less supported file format clusters, indicating that most clusters have more than 5 tools associated. For a small part of the clusters there are still no or very few tools known in the database as being able to process the associated file formats (6 clusters supported by up to three software tools). Archiving formats, text processing and image file formats clusters with strong tool support (about 100 tools or more) are presented in Figure 5. In conclusion, the application specific and not standardized formats are supported by a lower number of software tools according to the current version of the knowledge base.

4.2 Classification of related file formats

An existing categorization of file format types was used to verify the hypothesis that the formats with similar textual descriptions are related to each other (i.e. using alternative representations or encodings of the same type of data, allowing data conversion from one format to the other, etc.). The *List of file formats* available in Wikipedia presents the assignment of file format extensions to their types, which

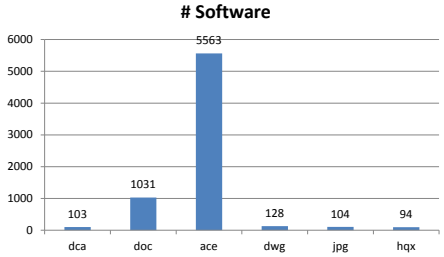

Figure 5: Clusters with strong software support.

Measure	Value
Number of file format clusters	30
Avg. file formats per cluster	13.6
Avg. format types per cluster	2.77
Avg. file formats of dominant type	21.09%
Avg. not classified file formats per cluster	64.26%
Avg. clusters with less than 2 categories	73.3%

Table 1: Statistics regarding the distribution of format types in related file formats (clusters)

are organized within a hierarchical structure [2]. In the first instance we used the Open Refine [3] tool for transforming the html representation of the categorization hierarchy to an appropriate taxonomy using the SKOS format. The later is available for download[4] from the FFMA server. The hierarchical taxonomy has the advantage of grouping categories of file formats that share certain commonalities (e.g. Raster Graphics and Vector Graphics are different possibilities of encoding Graphics content). The statistics on the type classifications of file formats over all clusters is presented in Table 1). There was accounted an average of 13.6 members per cluster and an average of 2.77 assigned format types (see also Figure 2 for cluster size distribution). The results presented here are highly influenced by the lack of categorization information, for 64% of the file extensions (available in the FFMA knowledge base) no file type assignment was found in the Wikipedia article. Under these circumstances about 21% of formats belonged to the *dominant* category and less than 15% was assigned to other categories. **Discussion.** The preliminary experimental results presented within this paper demonstrate the feasibility of the proposed approach. Anyway, no fine tuning of the clustering algorithm was performed, and no adjustments of the user generated taxonomy of file format types was made. Even so, the statistical analysis of the file formats presented in the Table 1 confirm our hypothesis that related file formats can be automatically identified using their descriptions (i.e. the average format types per cluster is 2.77, and 73% of the clusters have at most 2 categories). Still, this is not a strong evidence given the high rate of not categorized file formats. In time, we expect that more categorizations become available and the rate of formats of the dominant type to be significantly increased, even if the diversification of the format types per cluster might increase slightly. As future work we plan to significantly increase the rate of file format categorizations by taking in account more information sources like DBPedia genre, FileInfo classification [5], Yago formats [6], which will require spending significant efforts on ontology mapping purposes.

5. CONCLUSIONS

In this paper we present the enhancements added to the knowledge base of the file format metadata aggregator service. Artificial intelligence techniques are employed for identification of related file formats and to discover additional software tools that might be able to perform content migration between these formats. The preliminary evaluation demonstrates the feasibility of identifying similar formats basing on the textual descriptions acquired from the linked open data repositories. As future work we plan to use additional knowledge sources (e.g. vendor's web sites, further domain specific knowledge bases) for extending the knowledge related to the software tools, vendors and their relationship to the existing file formats.

6. REFERENCES

[1] U. C. Center. Unified digital format registry (udfr) - final report. Technical Report 2012-07-02, California Digital Library, University of California, 2012.

[2] R. Graf and S. Gordea. Aggregating a knowledge base of file formats from linked open data. In *iPress 2012*, pages 293–294, 2012.

[3] R. Graf and S. Gordea. A risk analysis of file formats for preservation planning. In *iPress 2013*, page to appear, 2013.

[4] H. Kulovits, C. Becker, M. Kraxner, F. Motlik, K. Stadler, and A. Rauber. Plato: A preservation planning tool integrating preservation action services. *LNCS - Research and Advanced Technology for Digital Libraries*, 5173:413–414, 2008.

[5] R. K. T. R. E. S. P. T. Orit Edelstein, Michael Factor. Evolving domains, problems and solutions for long term digital preservation. *iPRES 2011 - 8th International Converence on Preservation of Digital Objects*, 2011.

[6] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2:169–194, 1998. 10.1023/A:1009745219419.

[7] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[8] D. Tarrant and L. Carr. Lds3: applying digital preservation principals to linked data systems. In *Ninth International Conference on Digital Preservation (iPres2012)*, 2012.

[9] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. In *iPres2009: The Sixth International Conference on Preservation of Digital Objects*, June 2009. Event Dates: October 5th and 6th, 2009.

[2] see `http://en.wikipedia.org/wiki/List_of_file_formats`

[3] see `http://blog.semantic-web.at/2011/02/17/transforming-spreadsheets-into-skos-with-google-refine/`

[4] `http://ffma.ait.ac.at/taxonomies/FileFormatTypes`

[5] `http://www.fileinfo.com/`

[6] `http://dbpedia.org/class/yago/Format106636806`