# Aggregating a Knowledge Base of File Formats from Linked Open Data

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 semi-automatic aggregation of knowledge on computer file formats used to support planning for long term preservation. Our goal is to create a solid knowledge base from linked open data repositories which represents the fundament of the DiPRec recommender system. The ontology mapping approach is employed for collecting the information and integrating it in a common domain model. Furthermore, we employ expert rules for inferring explicit knowledge on the nature and preservation friendliness of the file formats.

## Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Systems issues; H.3.5 [**On-line Information Services**]: Data sharing

## 1. INTRODUCTION
The core of preservation planning is represented by the file formats used for encoding the digital information. Currently, the information about the file formats lacks a unified well-formed representation in LOD repositories and is only partially available in domain specific knowledge bases (i.e. PRONOM). The activities related to the preservation of digital content 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.

The linked open data (LOD) initiative defines best practices for publishing structured data in the Web using a well-defined and queryable format [3]. By linking together and inferring knowledge from different publicly available data repositories (i.e. Freebase, DBPedia, PRONOM) we aim at building a better, more complete characterization of available file formats. In this paper we present the File Format Metadata Aggregator (FFMA) service which implements the proposed approach for building a solid knowledge base supporting digital preservation planning and enactment. FFMA represents the core of the Digital Preservation Recommender (DiPRec) introduced in earlier paper by the authors [1]. The main contributions of this paper consist in: a) proposing and evaluating the approach based on ontology mapping for integrating digital preservation related information from the web; b) using AI models for inferring domain specific knowledge and for analyzing the preservation friendliness of the file formats basing on the expert models and computation of preservation risk scores.

## 2. KNOWLEDGE BASE AGGREGATION
One of the main concerns in the design of the FFMA service is the mapping of the semantics between LOD repositories and FFMA domain model. There are two alternatives for mapping file format ontologies: a) by employing ontology matching tools or b) by doing it manually [2]. For development of the FFMA service we chose to perform manual mapping (Fig. 1), due to reduce size of the domain model and the complexity and heterogeneity of the Freebase and DBpedia ontologies.

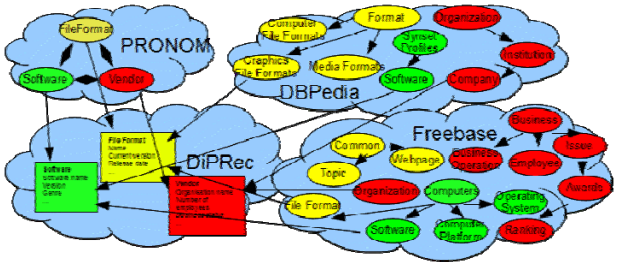


**Figure 1. Relationship between data representations.**

The underlying domain model consists of three core concepts represented by the File Formats, Software and Vendors. The properties associated to these objects are common for the LOD and PRONOM repositories. The FFMA domain model is aligned with the PRONOM one, which is a reference in the digital preservation domain. Since PRONOM data is not enough documented to cover all computer file formats, and their description is not rich enough for supporting reasoning and recommendations, we collect additional information from LOD repositories and aggregate it in a single homogeneous property based representation in the FFMA knowledge base (Figure 2).

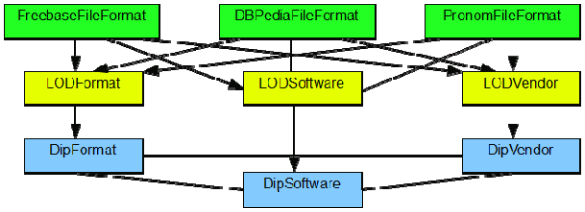


**Figure 2. FFMA domain object model overview.**

The *FileFormat classes store an index of the individual file format available in each of the external repositories, and they are used for crawling the LOD repositories for relevant information which is stored in LOD* objects. This data is cleaned from

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. iPRESS2012, Oct 1-5, 2011, Toronto, ON, Canada. Copyright 2012, Digital Curation Institute, iSchool, University of Toronto.

duplications or ambiguities and integrated in the internal representation of file format descriptions which is stored in the DipFormat, DipSoftware and DipVendor classes.

The Domain Knowledge Aggregation is based on the risk analysis model which is in charge of evaluating the information aggregated in the previous step and computing the risk scores over different digital preservation dimensions (e.g. provenance, exploitation context, web compatibility, etc.). A cost based model used for computing the risk scores is designed to provide a simple yet powerful mechanism for definition of expert rules, metrics and classifications used for computing recommendations in DiPRec. A more detailed description and examples on knowledge aggregation process can be found in [1].

## 3. EVALUATION

The aim of the experimental evaluation is to demonstrate the improvements provided by the proposed approach over the domain specific knowledge base represented by PRONOM. Apart from crawling the information basing on ontology mapping solution we also perform data cleaning in order to remove duplicates and ambiguous information.

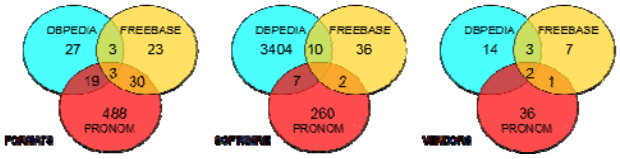


**Figure 3. The distribution of objects in LOD repositories.**

One of the possible practical user scenarios for FFMA system is the search of software solutions available for creation of the migration plans. The main goal of this scenario is to retrieve rich information on file formats, software and vendors from LOD repositories which allows evaluating the preservation friendliness of software formats.

In experiment we verified our hypothesis that information extraction from additional sources will significantly increase the amount of information available in PRONOM technical registry. The information extraction started with PRONOM (540 formats, 269 software and 39 vendors) was significantly enriched by DBPedia data (52 formats, 3421 software and 19 vendors) and concluded data retrieval with the Freebase (59 formats, 48 software and 13 vendors). In conclusion the FFMA knowledge base stores with ~10% more file formats, about 13 times more software and with 60% more vendors than PRONOM (see Fig. 3). Table 1 demonstrates a significant improvement of the aggregated information broken down to the sample file formats regarding additional knowledge about format versions, software and vendors. E.g. for “GIF” format FFMA comprises the description of 4 of its versions, 6 software tools and 2 vendors more than PRONOM. The multiple data entries in one LOD repository (e.g. two entries for "JPG" format in DBPedia) could be explained either with different versions of the same format or with slightly different names used for the same file format (i.e. identified by same extensions). Given the results presented above, we can demonstrate an important gain when aggregating knowledge from LOD repositories. Moreover, these repositories integrate data from public sources (e.g. like Wikipedia, Stock Market value for Software vendors, Websites of ISO/IETF standards, etc.) which is expected to be grow in time with the support of cross domain information sharing within the given communities.

**Table 1. Extracted file format values count in DiPRec classes.**

| Format | Versions | | Software | | Vendors | |
|---|---|---|---|---|---|---|
| | PR | FFMA | PR | FFMA | PR | FFMA |
| TIF | 9 | 19 | 0 | 134 | 0 | 1 |
| PDF | 17 | 33 | 14 | 30 | 5 | 6 |
| PNG | 3 | 7 | 13 | 28 | 4 | 5 |
| GIF | 2 | 6 | 13 | 19 | 4 | 6 |
| JPG | 9 | 12 | 13 | 16 | 4 | 5 |

## 4. CONCLUSIONS

Within this paper we presented the file format metadata aggregation service which builds a knowledge base with rich descriptions of computer file formats. The service uses semiautomatic information extraction from the LOD repositories, analyzes it and aggregates knowledge that facilitates decision making for preservation planning.

An important contribution of this paper is the usage of the ontology mapping approach for collecting data from LOD repositories. The evaluation of preservation friendliness is based on risk scores computed with the help of expert models. This allows automatic retrieval of rich, up to date knowledge on file formats, reducing so the setup and maintenance costs for the digital preservation expert systems (e.g. DiPRec).

As future work we plan to use additional knowledge sources (e.g. vendor's web sites, further knowledge bases) for extending the knowledge related to the software tools, vendors and their relationship to the existing file formats (which are often missing/incomplete in each of the named repositories). In the same time, we might consider to enhance the modules used for knowledge extraction for inferring further explicit knowledge (e.g. clustering by groups of file formats like text, graphical formats, video, audio file formats, etc.).

## 5. ACKNOWLEDGMENTS

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

## 6. REFERENCES

[1] Gordea, S., Lindley, A., and Graf, R. 2011. Computing recommendations for long term data accessibility basing on open knowledge and linked data. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 2011), 795-825.

[2] Jain, P., Yeh, P., Verma, K., Vasquez, R., Damova, M., Hitzler, P., and Sheth, A. 2011. Contextual ontology alignment of lod with an upper ontology: A case study with proton. In *G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, editors*, The Semantic Web: Research and Applications, vol. 6643 of LNCS. Springer Berlin, Heidelberg, 80-92.

[3] Kurt, B., Colin, E., Praveen, P., Tim, S., and Jamie, T. 2008. 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*, New York, NY, USA, 1247-1249.