Computing Recommendations for Long Term Data Accessibility basing on Open Knowledge and Linked Data

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

Andrew Lindley
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
andrew.lindley@ait.ac.at

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

ABSTRACT

Digital access to our cultural heritage assets was facilitated through the rapid development of the digitization process and online publishing initiatives as Europeana or the Google books project. As Galleries, Libraries, Archiving institutions and Museums (GLAM) created digital representations of their masterpieces new concerns arise regarding the long-term accessibility of digitized and digitally born content. Repository managers of institutions need to take well documented decisions with regard to which digital object representations to use for archiving or long term access to their valuable collections. The digital preservation recommender system presented within this paper aims at reducing the complexity in the process of decision making by providing support for classification and the preservation risk analysis of digital objects. Technical information which is available as linked data in open knowledge sources facilitates the construction of the DiPRec's recommender knowledge base. This paper presents the DiPRec recommender system, a community approach on how to achieve the generation of well founded and trusted recommendations through open linked data and inferred knowledge in the domain of long-term information preservation for GLAM institutions.

Categories and Subject Descriptors

H.3.7 [**Information Systems Applications**]: Digital Libraries; M.8 [**Knowledge Management**]: Knowledge Reuse

General Terms

Digital preservation, Recommender systems

Keywords

Knowledge based recommender, open recommendations, linked open data, preservation planning

1. INTRODUCTION

Knowledge based recommender systems (KBRs) as natural followers of expert systems are nowadays used for supporting the decision making process in multiple application areas as: e-commerce, financial services, tourism, etc. One of the most important challenges of KBRs is the construction of their underlying knowledge base. This is typically composed by sets of factual knowledge, i.e. information describing the application's domain and business rules. Both together enable the drawing of conclusions and support the decisions making process when analyzing the utility of a specific item in a given context as for example, analyzing the effectiveness of digitizing and publishing Mircea Eliade's book "History of Religious Ideas" within Google books.

Even though the world wide web has turned out to be the largest knowledge base, information published lacks an unified well-formed representation and mainly is intended for human readers. The Linked Open Data (LOD)[1] and Open Knowledge[2] initiatives address these weaknesses by describing a method on how to provide structured data in a well-defined and queriable format. By linking together and inferring properties of different independent and publically available information sources like FreeBase[3], DbPedia[4] and Pronom [5] within the specific context of a digital preservation scenario we shortcut the well known challenge of KBRs, the knowledge acquisition bottleneck.

In this paper we present our work carried out in the context of the Assets[6] project with the aim of preparing the ground for digital preservation within Europeana[7]. The Europeana portal serves as a central point for the large public to easily explore and research European cultural and scientific heritage online. It aggregates and collects data on digital resources from galleries, libraries, archives and museums accross Europe and by now manages about 19 million object descriptions collected from more than 15 hundred institutions. Within this very heterogeneous context it is easily understandable that digital objects are encoded in very heterogeneous file formats and versions throughout various different hardware and software content repository systems. Depending on the underlying use case it is likely that mul-

[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/
[6] http://www.assets4europeana.eu/
[7] http://www.europeana.eu/portal/

tiple representations of the same 'physical' object exist at a time. For example in most cases it is useful to provide access copies on demand which are easily distribuatable via the web while the master record needs to adhere to different requirements as for example the institution's long-term scenario and preservation policy.

A key topic in preservation planning is the file formats used for encoding the digital information. The Pronom Unique Identifiers (PUIDs) registry provides persistent, unique and unambiguous identifiers for file formats and therefore takes a fundamental role in the process of managing electronic records. Currently it lists information on about 820 different PUIDs. While some of the formats are properly documented, open-source and well supported, others may be outdated, redeemed by software vendors and no longer functional in modern operating systems. As always the the binary file's dependencies on the underlying platform, its configuration (codecs, plugins, etc.) as well as the rendering software are responsible on generating a concrete user performance, it is vital to have a solid understanding on all of them. This process is costly and requires a high degree of engineering expertise. Many of the GLAM institutions already outsource IT related activities and don't have the resources to keep track of the required level of complexity in house.

The Digital Preservation Recommender (DiPRec) system addresses the topics of 'preservation watch' and 'preservation policy recommendation'. It proposes a solution in the domain of digital long-term preservation for making documented recommendations based on risk scores, while the underlying knowledge base is built through a linked data approach. Information from FreeBase, DbPedia and Pronom in the areas of file formats, file conversions tools, hardware and software vendors is taken into account. The main contribution of this paper consists in the integration of open (general or domain specific) data when constructing knowledge based recommendations. The "knowledge acquisition bottleneck" and the high costs of setting up and maintaining KBRs are still an impediment for extensively adoption by the industry. Recommendations provided by DiPRec are meant to support GLAM institutions across Europe in the process of analyzing their digital assets. The technical foundation and the explanation of the DiPRec recommendations are computed on top of shared and collaboratively built data sources, trust in the area of LOD and digital preservation is a key issue which has been left out for this paper due to simplicity.

The novelty of our work consists in combining expert tools (as File, Droid or Fido) and automated object identification processes, with structured information (e.g. technical information on file formats) from open data repositories. This information is use for infering new knowledge, calculate preservation risks and finally for computing recommendations on preservation actions in the domain of digital long-term preservation. We present the rationale used for the construction of the DiPRec recommender by presenting concrete examples of a given content analysis which was provided for the Assets project. The rest of the paper is organized as follows; in Section 2 we present related work carried out on recommender systems and in the field of digital preservation. Section 3 highlights the architecture of DiPRec by comparing it against the construction of classical KBRs. The functionality provided by our system is explained in detail through a concrete example on the TIFF file format. The evaluation of our approach is presented in Section 4 by analyzing the digital collections of the Assets project. This is followed in the last Section of the paper (nr. 5) by the summarization of the concluding remarks for our work.

2. RELATED WORK

Knowledge Based Recommender systems gained broad popularity in e-commerce and e-tourism [7, 11, 24, 19] applications supporting customers in their decision making processes. The two most popular use cases are guidance through large and complex product offers (e.g. trip organization, feature selection of technical equipment) as well as accompanying the process of high cost decision making (e.g. financial investments). When designing the DiPRec recommender we took into consideration the Advisor Suite [12] and Planets Testbed infrastructure [16]. The main component of the Advisor Suite is a multipurpose workbench which offers support and advanced graphical user interfaces for constructing knowledge based recommenders. Advisor Suite features include the import of product catalogues, visual editing of a recommendation workflow and the generation of a runtime environment. The Planets[8] project focused on constructing practical services and tools for establishing empirical evidence in the process of informed decision making in the area of digital long-term preservation. A major achievement was the definition of basic nouns and verbs for core preservation operations. This allows to easily combine and swap tools within a preservation workflow and lead to a number of over fifty preservation services. Available services were deployed and tested within the Planets Testbed [22], a uniform environment for experimentation under well-defined and controlled surroundings. It provides automated quality assurance support for tools like DROID[9], JHOVE[10] and the eXtensible Characterisation Languages[11][5].

A key topic in preservation planning is the process of evaluating objectives under the limitation of well-known constraints. A state of the art report on technical requirements and standards as well as available tools to support the analysis and planning of preservation actions is given in [2]. Strodl et al. present the Planets preservation planning methodology Plato[12] by an empirical evaluation of image scenarios [21] and demonstrate specific cases of recommendations for image content in four major National Libraries in Europe[4]. After eliciting information regarding the preservation scenario (user requirements) the Plato tool is able to recommend specific preservation actions [3] for a given scenario. The tool was specifically designed to work on samples of the underlying data set and therefore is able to make use of XCL or similar tools for automated quality assurance and semi-automated evaluation of objectives. In contrast to these scenario evaluations, DiPRec aims at collecting information on a broader range from open linked data registries and dynamic knowledge sources. It can evaluate more general, even 'non-technical' objectives (e.g. what is the risk that no software vendor will support old formats like Word

[8] http://www.planets-project.eu/
[9] http://droid.sourceforge.net/
[10] http://hul.harvard.edu/jhove/
[11] http://planetarium.hki.uni-koeln.de/public/XCL/
[12] http://www.ifs.tuwien.ac.at/dp/plato/intro.html

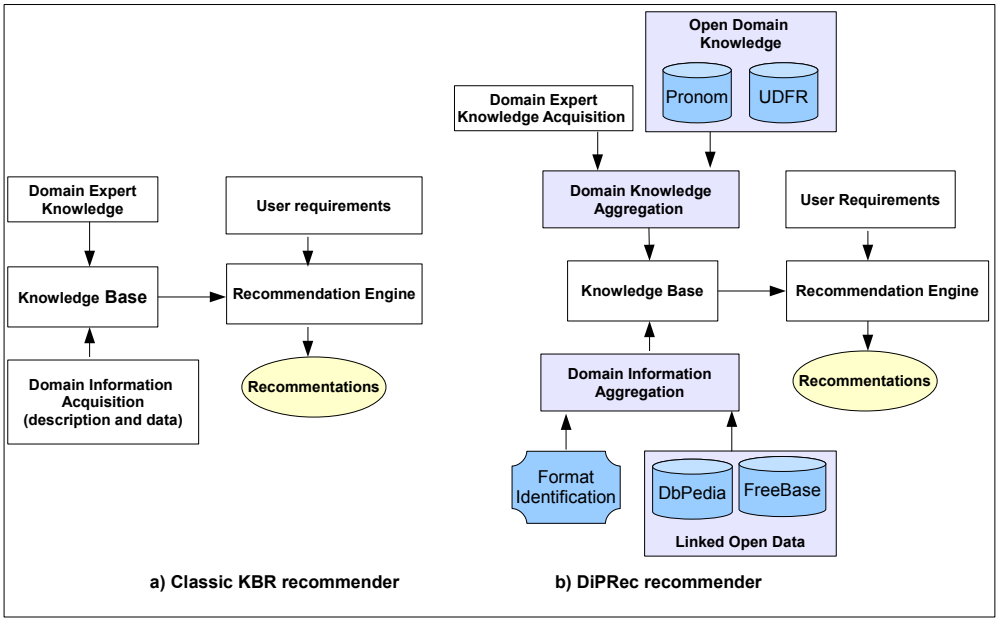

Figure 1: A comparison of regular KBRs and DiPRec recommender processes

3 documents?). This is a significant improvement over the Plato tool where all this information needs to provided by domain experts.

The Scape[13] project is one of the major current initiatives [18] which is partially funded by the European Union's FP7 on institutional preservation requirements. The project addresses besides the issues of scalable preservation and quality-assured preservation workflows also the topic of policy-based preservation planning and watch.

The paradigms of semantic Web and linked open data [6] transform the web from a pool of information into a valuable knowledge source of data according to the definitions of a knowledge management theory [17]. The exploitation of linked data as knowledge source for recommender system started as research topic in the last few years and was first applied to improve case-based and collaborative filtering recommenders [10, 9, 20]. In [20] the authors present the Talis Aspire system which is able to assists educational staff in picking educational web resources. The employment of linked data in collaborative filtering and case-based reasoning was explored by Heitmann and Hayes in [9] and [10].

3. SYSTEM OVERVIEW

Typically the creation of classic knowledge based recommender systems consists of three main tasks. Dealing with the collection of detailed descriptions of products offers is followed by the process of constructing a recommendation knowledge base (see section 3.2). At runtime user requirements elicitation takes place and recommendations are computed based on the underlying recommendation knowledge base and the items that match the given user requirements. DiPRec follows the same process but improves the way the knowledge base is built in order to reduce the efforts spent on domain knowledge acquisition. This is especially relevant for being exploited in GLAM preservation scenarios, where the underlying knowledge base contains broader information than the domain specific KBRs. Within the DiPRec recommender the Domain Information Aggregation module is responsible for collecting file format related information (e.g. formats, vendors, applications, etc.) from the open knowledge bases Pronom, DBPedia and Freebase. Furthermore the Domain Knowledge Aggregation module combines the outcome of a risk analysis process with the knowledge manually provided by domain experts. Figure 1 compares the process used by regular KBRs and the one presented by DiPRec which enhances the process of building the underlying knowledge base. In the following sections we present extended details on how the knowledge base of DiPRrec is built by using as example the Tagged Image File Format (TIFF).

The TIFF format is still very popular among the publishing industry, as it is a very adaptable file format although it did not have a major update since 1992. It was originally created by Aldus and since 2009 it is now under control of Adobe Systems. There are a number of extensions available (e.g. TIFF/IT, TIFF-FX) which have been based on the TIFF 6.0 specification, but not all of them are broadly used. A standard and broadly accepted approach in the archiving world is the migration of TIFF encoded content to the JPEG2000 format. In [4, 2] one can find the context in which several content providers took the decision to perform this kind of content migration. However within these scenarios, the context evaluation and the recommendation were computed by domain experts and by expert systems.

The DiPRec system, on the one hand applies to the approach of well-documented and trackable decision making, and at the same time it uses a semi-automatic approach on domain knowledge acquisition. This reduces the human effort invested by domain experts when providing reservation recommendations, reduces the financial efforts invested in the context evaluations, and in the same time is able to offer good quality recommendations.

[13] http://www.scape-project.eu/

FILE FORMAT DESCRIPTION	
Format Name	Tagged Image File Format (P), Tagged_Image_File_Format (D), Tagged_Image_File_Format(F)
Pronom Id	fmt/10 (P)
Mime Type	/media_type/image/tiff-fx, /media_type/image/tiff (F), image/tiff(P)
File Extensions	.tiff, .tif (D)
Current Version	6 (P)
Current Version Release Date	03 Jun 1992 (P)
Software License	Proprietary software (D)
Software	QuickView Plus, Acrobat, AutoCAD, CorelDraw, Freemaker, GoLive, Illustrator, Photoshop, Powerpoint (P), SimpleText, Seashore, Imagine (D)
Software Homepage	`http://adobe.com/photoshop(D)`
Operating System	PC, Mac OS X, Microsoft Windows (D)
Genre	Image (Raster) (P), Image file format (I), SimpleText - Text editor, Adobe Photoshop - Raster graphics editor (D)
Open Format	none (P)
Standards	ISO 12639:2004 (W)
Vendors	Aldus, Adobe Systems, Apple Computer, now Apple Inc., Microsoft (D), Adobe Systems Incorporated (P), Aldus Corporation (P)
VENDOR DESCRIPTION	
Organization Name	Adobe Systems
Country	United States (P)
Foundation date	Dec 1982 (F)
Number of Employees	6068 (Jan 2007), 8660 (2009)(F), 9,117 (2010)(W)
Revenue	3,579,890,000 US$ (Nov 28, 2008) (F)
Homepage	`http://adobe.com/photoshop(F)`

Table 1: File format and vendor description. (Information sources P = Pronom, D = DBPedia, F = Freebase, W = Wikipedia)

3.1 Domain Information Aggregation

Differently to the e-commerce domain where KBRs import detailed item descriptions from product catalogs there is no such catalog for computer file formats. The Unified Digital Format Registry (UDFR)[14] project was started in 2009 by a group of Universities and GLAM institutions with the aim of building a single, shared technical registry for file formats based on a semantic web and linked data approach. The project is based on the Pronom database which provides basic information about a large number of file formats and will be extended by data on migration pathways and available software/tools. The registry should be available from the beginning of 2012. As Pronom data is not rich enough to build a recommendation and reasoning mechanism for preservation scenarios of file formats on top, we collect additional information sources and aggregate them into a single homogeneous property representation in the recommender's knowledge base. DiPRec uses two types of operations for aggregating domain information:

- data unification: the data representation retrieved from different knowledge bases is unified and combined under the DiPRecs property model definition. For example, the number of software tools supporting a given file format is calculated over different data sources. 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.

- property composition: more abstract properties which require a hierarchical composition are computed by aggregating basic properties by weighted numbers. The model on property definition is meant to be kept very simple. For example "supported by major vendors" will check if at least one of the software companies is considered to fulfill this requirement by combining the properties like "NUMBER OF EMPLOYEES", "VENDOR REVENUE"). See Table 2.

When aggregating domain information we are interrogating external knowledge sources like DbPedia and Freebase which manage huge amounts of linked open data triples. This allows us to extract fragmental descriptions on file formats, software applications and vendors supporting given file formats (see Table 1). DbPedia allows to post sophisticated queries using SPARQL query and OWL ontology languages [13] for retrieving data available in Wikipedia. Freebase [15] is a practical, scalable semantic database for structured knowledge and is mainly composed and maintained by community members. Public read/write access to Freebase is allowed through an graph-based query API using the Metaweb Query Language (MQL) [6]. PRONOM data is released as linked open data and is accessible through a public SPARQL endpoint.

AGGREGATED PROPERTIES	
File format related	
Is supported by major software vendors?	yes
Is an open file format?	no
Is widely supported by current web browsers?	yes
Which versions officially supported by vendor?	6.0
Which versions are frequently used?	6.0
Image file compression supported?	yes
Preservation related metadata	
Is creator information available?	yes/no
Is publisher information available?	yes/no
Is digital rights information available?	yes/no
Is file migration allowed?	yes/no
Object creation date?	datetime
Is an object preview available?	URL

Table 2: Sample compound properties.

[14] `http://www.udfr.org/`

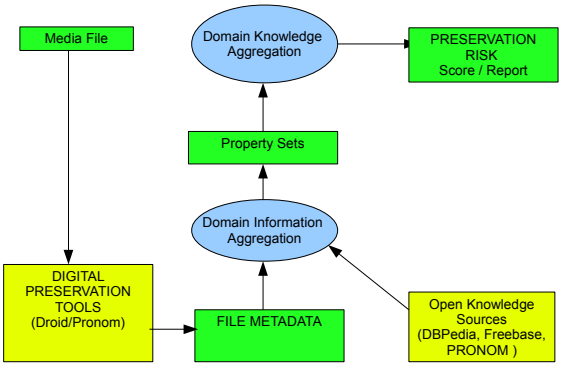

Figure 2: Domain knowledge aggregation process.

3.2 Domain Knowledge Aggregation

Pronom as presented before is a viable resource for anyone requiring impartial and definitive information about the file formats, software products and other related data. Extremely valuable to the DiPRec recommender is the information related to the file conversion tools based on a given PUID. Therefore we employ the Droid [15] characterization service for automatically extracting technical metadata and identifying file formats from physical media files. This metadata is then used in conjunction within the domain knowledge aggregation process presented in the Fig. 2

The risk analysis module is in charge of evaluating information previously aggregated in the DiPRec knowledge base for a given record at hand over following (exemplary) dimensions of digital preservation:

- Web accessibility: Dissemination copies are published and accessible on e.g. the content provider's web portal. There should be previews of objects (e.g. thumbnails for images, video summaries, short intro for audio files) and 'rich' object descriptions to increase their visibility and retrieval. The chosen file representations should render in the latest browsers without plugin support and cope with modern features (e.g. pseudo streaming, progressive image display, HTML5, X3D, etc). Content is made available through different exploitation channels.

- Archiving and costs: The decision of following a specific institutional preservation policy for a given technology is heavily influenced by given hardware and budget constraints. Future exploitations on the costs for content exploitation need to be predicted and taken into account.

Other scenarios may include:

- Provenance metadata
- Data exchange and collaborative data enrichment
- Publishing and digital rights management

The definition of preservation dimensions is not orthogonal and therefore certain properties might be involved more than once when computing different risk score. Due to management and maintenance reasons properties are also grouped by sets and 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 value of the overall risk score for a given collection of objects is computed as a weighted sum over all digital preservation dimensions:

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

Where R_i represents the preservation risk computed over the dimension i. ps represents the index of the current property set within all sets associated to the dimension i. The $w_{(ps,i)}$ is the weight of the contribution of the property set ps to the 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 $d(p, PFV(p))$.

3.3 User requirements elicitation

DiPRec is designed to work as a multi-purpose digital preservation support tool which can be used in various scenarios by different types of customers. For examples the tool may support content providers in analyzing the 'preservation friendliness' of their infrastructure, their archiving solutions or the visibility of their artifacts published in the Europeana portal. Recommendations are always to be seen in the context in which the digital objects are used. Within the scope of the Assets project there is the common interest to offer public access to digital assets through the Europeana portal (i.e. web discovery), to provide advance search functionality (i.e. description richness and preservation of provenance information) as well as the topic of the data archiving dimension.

As a result of the requirements elicitation process user profiles are created. A set of multiple choice questions is used to distinguish the relevant dimensions of available preservation objectives. According to different levels of complexity, role and required domain knowledge the system offers a subset of questions which are well understood and the best available choice for a user to express his needs. Fig. 3 presents sample workflow which could be used to determine a given user profile. For example a private user (ut = *private person*) with a solid level of IT knowledge (itk = *expert*) will be asked about preferred encodings and compression types of the digital content, while others would define attributes about storage limitations and upload samples of a given collection.

3.4 Recommendation computation

Differently to classic KBRs where the application's scope is very well delimited in terms of selecting the best matching items in a list of known possibilities, the DiPRec system relies on expressing an institutional preservation context in form of user requirements that are combined with the knowledge acquired about the long term accessibility threatening. We employ tools to evaluate the content of a given collection from a technical point of view and to generate fine grained preservation risk scores. When records are identified to have vulnerabilities on certain preservation dimensions a

[15] `http://sourceforge.net/projects/droid/`

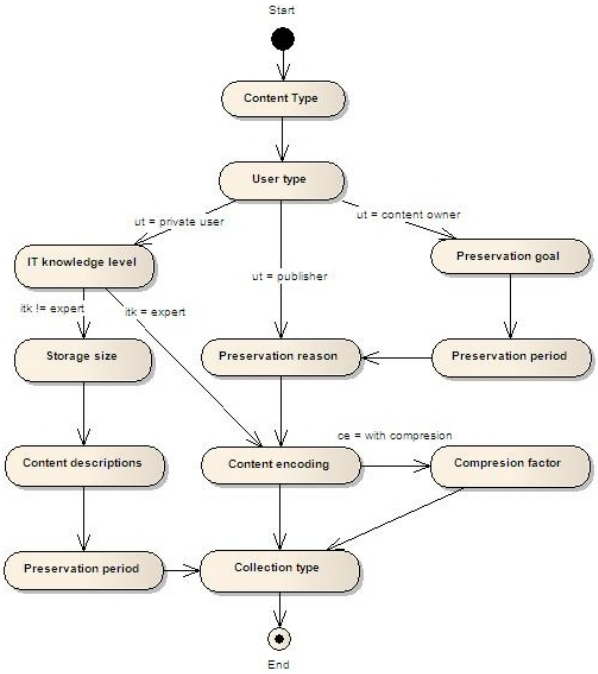

Figure 3: User requirements elicitation workflow

rule based engine as JBoss Drools[16] is used to propose appropriate preservation actions. The set of available business rules are defined by domain experts in form of simple IF-THEN-ELSE rules. These rules are neither complete nor meant to be non-overlapping. Unified tool access for processing executable preservation plans is provided through the Assets preservation normalisation framework which is able to invoke the tools with exactly defined settings and parameter configurations.

IF ($r_{ac} > 0.5$ AND i_a == true AND i_{wa} == true AND $open\_format$ == FALSE)
THEN migrate($preservation\_format$)

IF ($content\_type$ == IMAGE)
THEN preservation_format = (JPEG/2000:1, TIFF/6:0.8)

IF ($file\_format$ == TIFF/5 AND $preservation\_format$ == JPEG/2000)

$$THEN migration\_tool = IMAGE\_MAGICK \quad (2)$$

The preservation recommendations are computed using the constraint solving problems (CSP) theory [8, 11]. Constraints are defined within the preservation actions knowledge base, the CSP context is defined by user profiles and the preservation risks are identified for the given data collection. The recommendations are represented in form of preservation actions. For example, the set of business rules defined above combined with a user profile indicating interest in the dimension of archiving and web accessibility will lead to the following recommendation when analyzing a collection of images in TIFF format:
$migrate(TIFF/5, JPEG/2000, IMAGE\_MAGICK)$
In free text translation, the recommendation will suggest the migration of the files available in $TIFF/5$ format to $JPEG/2000$ by using the $IMAGE\_MAGICK$ software with standard settings.

[16]http://www.jboss.org/drools

4. EVALUATION

The evaluation of the first prototype of DiPRec was conducted within the scope of the Assets project. Ten partners of the project consortium provided metadata and binary content (10 collections with a total size of 516GB contained in 368067 media files) for supporting the development and testing of services developed within the scope of the project. The first step in the evaluation process was the identification of file formats, definition of property sets and the aggregation of the domain knowledge available in open knowledge bases on these file formats.

The Table 3 lists the distribution of file formats by content type. Even the experimental data was taken from a small number of content providers, we discovered a variety of 18 formats in 38 different versions used for encoding the digital content.

Content Type	File Format	# Versions	# Files
TEXT	TXT	1	4
TEXT	DOC	1	16
TEXT	XML	1	20101
TEXT	HTML	1	1205
IMAGE	JPG	8	323332
IMAGE	PSD	1	3
IMAGE	PNG	4	1228
IMAGE	BMP	2	141
IMAGE	GIF	2	1066
IMAGE	TIFF	4	4
IMAGE	PDF	16	25008
AUDIO	MP3	1	3634
VIDEO	FLV	1	9468
VIDEO	MPEG4	1	935
VIDEO	MPEG1	1	3074
VIDEO	MPEG3	1	3074
3D	PLY	1	50
3D	DAE	1	307

Table 3: Distribution of file formats in Assets collections.

The Digital Record Object Identification tool (DROID) version 5, signature file 45 was executed through the Assets preservation normalisation tool suite and was able to successfully identify file formats in 95 percent of the cases through its binary signature method except of the 3D model objects which have not yet been collected by Pronom. Appropriate information on all of the file formats was contained in DbPedia and Freebase and the domain knowledge acquisition process was completed by successfully computing the preservation risk analysis scores.

The second part of the evaluation consisted in computing recommendations for the given content. Therefore, we created a user profile for content providers that are interested in making their content accesible through Europeana. Within this context, the content providers manifest interest for the web accessibility digital preservation dimension.

The highest diversity of file formats was found in the image collections. The recommendation to migrate these files to the JPEG 2000 format didn not get a high priority

and will be performed within the next period of scheduled storage migration. The Image Magick tool was the recommended choise for performing this transformation action. The whole audio content available in Assets was provided in the mp3 format and no recommendation was made for transforming audio collections. The most restrictive constraints for web accessibility are defined for the video content. The pseudostreaming protocol is an advanced technological solution used for distributing information over the web. It allows the user to interact with the media-player and to quickly navigate within the content without the needed to download the entire media file. This protocol is supported by two file formats: flash video (FLV) and MPEG4 with H2.64 video encoding. It has native support in HTML5 and is used in HTML4 with an adequate browser plugin. A part of the Assets content is already available in FLV format and another part is available in MPEG1 or MPEG2. The DiPRec resulting recommendation is to migrate the content to FLV by using the ffmpeg [17] tool.

5. CONCLUSION

Within this paper we introduced the DiPRec recommender system, an expert support tool in the domain of digital long-term preservation for GLAMs. An important contribution of this papers is the exploitation of an open linked data approach for constructing the recommender's knowledge base built upon open registries as DbPedia and Pronom. Since the knowledge acquisition, aggregation and unification process is fully automated it is easy to upgrade the recommender's knowledge base.

We looked at preservation planning which is the process of specifying clearly defined and relevant trees of objectives in a defined preservation dimension and evaluating them within a given (institutional) context to generate well-documented decisions. DiPRec is able to advance the process with inferred community knowledge and reduces the degree of manual evaluation processes or require technical expertise in this process.

Am important concern related to the KBRs is the trust in the provided recommendations. This is especially relevant for the digital preservation domain where we deal with a large amount of multimedia material and the execution of the preservation actions is associated with considerable costs. Within this paper we did not examine the completeness, correctness and quality degree of the underlying data. We however argue that data from open knowledge bases like DbPedia or Freebase could protect from biases introduced by the economical interests of professional companies by its underlying community approach.

The tool has been designed by reusing our past experience in building knowledge based and case based recommender systems [23, 8] and combining it with the expertise of creation long-term preservation infrastructure and applications [14, 1]. Based on this work the Assets normalisation tool suite is able to automate the process of object identification and characterisation and therefore directly integrates within the property evaluation, risk analysis and recommendation process for a given record. We presented a first evaluation of digital content provided by national libraries and archives through the Assets project where the underlying concepts of the DiPRec approach were proven to work adequately.

[17] http://www.ffmpeg.org/

6. ACKNOWLEDGMENTS

This work was partially supported 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).

7. REFERENCES

[1] Aitken, B., Helwig, P., Jackson, A., Lindley, A., Nicchiarelli, E., Ross, S.: The planets testbed: Science for digital preservation. Code4Lib 1(3) (2008), http://journal.code4lib.org/articles/83

[2] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., Hofman, H.: Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. International Journal on Digital Libraries 10(4), 133–157 (2009), http://dblp.uni-trier.de/db/journals/jodl/jodl10.html#BeckerKGSRH09

[3] Becker, C., Kulovits, H., Rauber, A., Hofman, H.: Plato: a service-oriented decision support system for preservation planning. In: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries. pp. 367–370. ACM, New York (2008), http://publik.tuwien.ac.at/files/PubDat_170832.pdf, vortrag: 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2008), Pittsburgh, Pennsylvania; 2008-06-16 – 2008-06-20

[4] Becker, C., Rauber, A.: Four cases, three solutions: Preservation plans for images. Tech. rep., Vienna University of Technology, Vienna, Austria (April 2011)

[5] Becker, C., Rauber, A., Heydegger, V., Schnasse, J., Thaller, M.: A generic xml language for characterising objects to support digital preservation. In: SAC '08: Proceedings of the 2008 ACM symposium on Applied computing. pp. 402–406. ACM, New York, NY, USA (2008)

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

[7] Burke, R.D.: Hybrid web recommender systems. In: The Adaptive Web. pp. 377–408 (2007)

[8] Felfernig, A., Gordea, S.: Ai technologies supporting effective development processes for knowledge-based recommender applications. In: SEKE. pp. 372–379 (2005)

[9] Heitmann, B., Hayes, C.: C.: Using linked data to build open, collaborative recommender systems. In: In: AAAI Spring Symposium: Linked Data Meets Artificial IntelligenceŠ. (2010

[10] Heitmann, B., Hayes, C.: Enabling case-based reasoning on the web of data. In: The WebCBR Workshop on Reasoning from Experiences on the Web (2010)

[11] Jannach, D., Zanker, M., Fuchs, M.: Constraint-based recommendation in tourism: A multiperspective case study. J. of IT & Tourism 11(2), 139–155 (2009)

[12] Jannach, D., Zanker, M., Jessenitschnig, M., Seidler, O.: Developing a conversational travel advisor with advisor suite. In: ENTER'07. pp. 43–52 (2007)

[13] Jens, L., Jörg, S., Sören, A.: Discovering unknown connections -the dbpedia relationship finder. In: Proceedings of the 1st Conference on Social Semantic

Web (CSSW). vol. P-113, pp. 99–109. Gesellschaft für Informatik, Leipzig, Germany (2007)

[14] King, R., Schmidt, R., Jackson, A., Wilson, C., Steeg, F.: 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. vol. 5714/2009, pp. 425–428. Springer-Verlag (2009), http://dx.doi.org/10.1007/978-3-642-04346-8_50

[15] Kurt, B., Colin, E., Praveen, P., Tim, S., Jamie, T.: 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. pp. 1247–1249. ACM, New York, NY, USA (2008)

[16] Lindley, A., Jackson, A.N., Aitken, B.: A collaborative research environment for digital preservation - the planets testbed. Enabling Technologies, IEEE International Workshops on 0, 197–202 (2010)

[17] Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press (May 1995)

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

[19] Ricci, F., Werthner, H.: Case base querying for travel planning recommendation. Journal of IT & Tourism 4(3-4), 215–226 (2001), http://dblp.uni-trier.de/db/journals/jitt/jitt4.html#RicciW01

[20] Shabir, N., Clarke, C.: Using linked data as a basis for a learning resource recommendation system. In: 1st International Workshop on Semantic Web Applications for Learning and Teaching Support in Higher Education (SemHE'09) (September 2009), http://eprints.ecs.soton.ac.uk/18053/

[21] Strodl, S., Becker, C., Neumayer, R., Rauber, A.: How to choose a digital preservation strategy: evaluating a preservation planning procedure. In: JCDL '07: Proceedings of the 2007 conference on digital libraries. pp. 29–38. ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1255175.1255181

[22] Sven Schlarb, Edith Michaelar, M.K.A.L.B.A.S.R.A.J.: A case study on performing a complex file-format migration experiment using the planets testbed. IS&T Archiving Conference 7, 58–63 (2010)

[23] Zanker, M., Gordea, S., Jessenitschnig, M., Schnabl, M.: A hybrid similarity concept for browsing semi-structured product items. In: EC-Web. pp. 21–30 (2006)

[24] Zanker, M., Jessenitschnig, M., Jannach, D., Gordea, S.: Comparing recommendation strategies in a commercial context. IEEE Intelligent Systems 22(3), 69–73 (2007)