Automating Logical Preservation for Small Institutions with Hoppla

Stephan Strodl, Petar Petrov, Michael Greifeneder, and Andreas Rauber

Vienna University of Technology, Vienna, Austria
{strodl,petrov,greifeneder,rauber}@ifs.tuwien.ac.at

Abstract. Preserving digital information over the long term becomes increasing important for large number of institutions. The required expertise and limited tool support discourage especially small institutions from operating archives with digital preservation capabilities. Hoppla is an archiving solution that combines back-up and fully automated migration services for data collections in environments with limited expertise and resources for digital preservation. The system allows user-friendly handling of services and outsources digital preservation expertise. This paper presents the automated logical preservation process of the Hoppla archiving system in detail. It describes the recommendation process for appropriate preservation strategies via a web update service. A set of two real world case studies were conducted based on a first rules set focused on common office documents. The promising results sustain the novel approach of automating logical preservation by outsourcing expertise.

1 Introduction

Digital information form essential assets in the long run for an increasing number of small institutions. Not only legal obligation mandate archiving of digital objects, but the loss of data can lead to serious business problems, e.g. data containing intellectual property, know-how, expertise or business data.

The common practice is regular backup on storage devices be they external hard disks or DVDs. Most users do not know the exact specification or format of their digital objects. Neither are they aware of changes in technological environment and therefore which formats could still be rendered in 5, 10 or even 15 years. Although the bitstream preservation problem is not entirely solved, there exist many years of practical experience in the industry, with data being constantly migrated to current storage media types, and duplicate copies held to preserve bitstreams over years.

A much more pressing problem is logical preservation. The rendering of a bitstream depends on the environment of hardware platforms, operating systems, software applications and data formats. Even small changes in this environment can cause problems in opening an object. Digital preservation is mainly driven by memory institutions like libraries, museums and archives, which have a focus on preserving scientific and cultural heritage, and dedicated resources available

M. Lalmas et al. (Eds.): ECDL 2010, LNCS 6273, pp. 124–135, 2010.

to care for their digital assets. Enterprises whose core business is not data curation are going to have an increased demand for knowledge and expertise in logical preservation solutions to keep their data accessible. Long-term preservation tools and services are developed for professional environments to be used by highly qualified employees in this area. In order to operate in more distant domains, automated systems and convenient ways to outsource digital preservation expertise are required.

The Hoppla Archiving System[1] [9] combines back-up and fully automated migration services for data collections in small office environments. The system allows user-friendly handling of services and outsources digital preservation expertise. Hoppla uses its migration capabilities to continuously migrate digital objects which are in danger of being obsolete and unaccessible in the near future to more stable formats. The knowledge how and under which circumstances to migrate a certain object is provided by experts. This information is distributed to every Hoppla client upon request via a central web service.

The major challenges of an automated archiving system are the decision-making ability and the error tolerance of the software system. In particular the migration process is highly error-prone and needs special automated error handling mechanism due to the limited competence of users of troubleshooting and decision making. A further challenge is the variety of formats in the collections of small intuitions. The archiving system need to provide migration pathways for a large number of formats that are at risk of becoming obsolete. This paper presents the automated logical preservation process of the Hoppla system in detail.

The remainder of this paper is organised as follows: Section 2 provides pointers to related initiatives and gives an overview of work previously done in this area. Section 3 describes the automated logical preservation process of the Hoppla system, followed by the results from a first set of case studies in Section 4. An outlook on future work and a final conclusion is presented in Section 5.

2 Related Work

The concept and the design of the Hoppla system is presented in [9]. A number of research initiatives have emerged in the last decade in the field of digital preservation, primarily memory institutions focusing on professional environments. The raising awareness for small institutions and SOHOs increases demand of practical solutions for users with less experience [2].

Existing open source digital repositories, such as Fedora Commons[2] and DSpace[3], are developed for large scale collections in professional archiving. These repositories provide a huge function range, but require considerable knowledge for configuration and usage. The overhead of function and configuration make

[1] `http://www.ifs.tuwien.ac.at/dp/hoppla`
[2] `http://www.fedora-commons.org`
[3] `http://www.dspace.org`

these systems unsuitable for institutions with limited knowledge in data management. The innate support of these systems for logical preservation is limited. Considerable effort of integration and development would be necessary to provide long term preservation functionality for a collection. Another repository such as the e-Depot [8], developed by KB and IBM focus on electronic publications and is also developed for use in professional settings.

The CRIB project [3] has developed a Service Oriented Architecture implementing migration support. The digital objects are transferred to a server infrastructure and migrated objects are returned. The actual migrations of the objects are executed on the server side. CRIB is integrated into the RODA repository[4]. Transferring complete data collection to an external web service is potentially inefficient for small institutions and raises privacy issues.

The Panic Project [5] developed a framework to dynamically discover preservation strategies with similarities to the Hoppla system. Panic uses semantic web technologies to make preservation software modules available as Web services. The system is designed for large-scale repositories that implement the required services invoker. Panic uses external web services with actual data similar to the CRIB project. The Hoppla web service on the other hand provides only preservation recommendations to the client system. The migrations of the data are executed on the client side without transferring private data via the internet. Moreover the Hoppla system includes the bit preservation of the data.

The PreScan system [7] automatically extracts embedded metadata from digital objects. The system scans objects on a hard disc and manages their metadata in an external repository that supports Semantic Web technologies. The metadata could be used to implement digital preservation support.

A range of projects are developing tools and components to support digital preservation. Format registries that can be used to determine required preservation actions are developed by UK National Archive's PRONOM project [10] and the Global Digital Format Registry (GDFR) [4]. The format identifier tool Droid[5], based on the Pronom registry, is used in the Hoppla system to determine objects format. In order to obtain additional metadata about objects, a number tools and services are developed. For example, JHove[6], developed by JSTOR and the Harvard University Library is used within Hoppla.

Research on technical preservation issues is focused on two dominant strategies, namely migration and emulation. The Council of Library and Information Resources (CLIR) presented different kinds of risks for a migration project [6]. Migration requires the repeated conversion of a digital object into more stable or current file format. Migration is a modification of the data and always incurs the risk of losing essential characteristics of the object [6]. Still the number of tools as well as the ease of applying migration makes it a very promising candidate for archiving in small institutions. Emulation, the second important preservation strategy aims at providing programs that mimic a certain environment. The

[4] http://roda.di.uminho.pt
[5] http://droid.sourceforge.net
[6] http://hul.harvard.edu/jhove

major disadvantage is that the emulator itself is a piece of software and has to be preserved over time. In order to keep the archiving system simple and easy to apply, we are currently not considering emulation as a preservation strategy for Hoppla.

3 The Hoppla Logical Preservation Process

The workflow and the design of the Hoppla system is presented in [9]. In this paper the automated logical preservation process of the Hoppla archiving system will be presented in detail. The basic idea is to provide preservation know-how to client users with limited expertise. Therefore the client sends information about the collection, the available migration tools and the user's requirements to a central update service (Figure 1). Based on the information appropriate preservation rules and tools for objects that are at risk of becoming obsolete are selected by the update service and send to the client. Migrations of the objects are executed according the provided rules on the client side. The process is described in detail in the following sections.

3.1 Collection, User and System Profiling

The first step in the process is to create a collection profile on the client side. The profile describes the technical characteristics of the collection to be preserved. The description includes the objects formats as well as more detailed description, for example the encoding of videos, size of the object, transparent layers of images, etc. This information is essential to select suitable migration strategies. The profile consists of aggregated metadata extracted from digital objects in the collection. The metadata are created within Hoppla by using format identification tools (e.g. DROID) and characterisation tools (e.g. JHOVE).

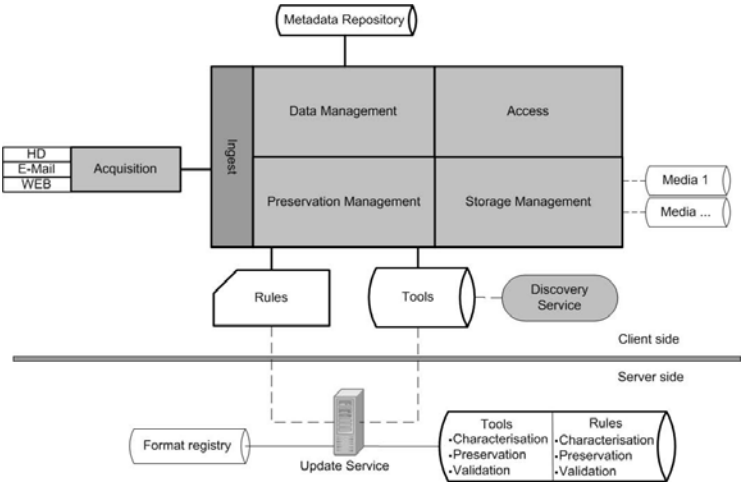

Fig. 1. Architecture of the Hoppla System

User's requirements on the digital data differ significant between user groups. These requirements and particularly future usage scenarios of the digital objects have substantial influence on the selection of preservation rules. For example a professional photographer has other future usage scenarios for images (enlarging, cutting, editing) than other users and therefore need other migration rules. In order to capture the preferences in a user-friendly way Hoppla has introduced a logical preservation level rating for four types of objects (text, audio, video and images). The rating ranges from zero (no preservation actions) to one (all available actions). According the rating preservation rules are selected. For a low rating only essential preservation actions for objects that are at imminent risk of becoming obsolete are performed. For a medium rating migrations are done for formats that are no longer in wide-spread use, outmoded (e.g. old versions of application) or the migration of proprietary format to open source formats. High ratings initiate additional multiple migration pathways for object formats to support different future usage scenarios e.g. migration of Microsoft Word objects to PDF/A, OpenOffice Format(ODT) and plain text format (preserving the layout, the editability and allowing computer-assisted processing). The selection of the preservation strategies is presented in detail in Section 3.3.

In addition to collection characteristics and user preferences the technical environment on the client side effects the selection of preservation rules. It is described in the system profile. The profile consists of technical characteristics (such as operating system, free storage size, etc.) and installed migration tools on the client system. A main challenge of an automated archiving system is the provision of useful migration services for large number of object formats. Hoppla supports two kinds of migration service, portable and external. Portable migration services are provided by the update service. The services are downloaded and dynamical integrated into the Hoppla client application. This mechanism works excellent for limited number of migration tools, for example tools implemented in Java. In order to provide migration support for a larger variety of formats Hoppla uses migration services installed on the client system (so called external services). A service to discover installed tools is described in the next Section 3.2.

3.2 Tool Discovery

In order to provide migration paths for a wide range of formats Hoppla uses migration services installed on the client host system. A discovery service helps to identify tools available on Linux or Windows systems. The current version of the discovery tool only supports tool that can be executed via the command line. The migration services to discover are defined in a XML configuration file, which can easily be updated through the web service. Examples for migration services are ImageMagick[7], OpenOffice[8], Mencoder[9].

[7] `http://www.imagemagick.org`
[8] `http://www.openoffice.org`
[9] `http://www.mplayerhq.hu`

For Linux the system executes the command and parses the outcome to check whether the tool is installed on the system and which version of the tool. Typical installation paths (e.g. /usr/bin/) can be specified to allow a more precise search. On Windows machines the system registry is searched for migration tools and their installation paths. The result of the discovery services is a list of installed migration tools including their installation path on the system. These tools can be used by the Hoppla system for migrations.

3.3 Selection of Preservation Strategies

One of the most challenging tasks in the preservation process is the selection of appropriate preservation strategies for a given collection. A profound and solid evaluation of different preservation alternatives is a time-consuming task and requires input by different domain experts (e.g. technic, curation, etc.). Work on preservation planning has been done for example with the planning tool Plato [1]. It was developed to support the evaluation of different potential preservation strategies against individual preservation requirements. The requirements are collected from the wide range of stakeholders and influence factors that have to be considered for a given setting. The Plato approach was designed to be used in professional settings by preservation expert. Due to the limited expertise and knowledge of the typical Hoppla user an expert preservation planning process (such as Plato) is not feasible and an alternative approach is required. The Hoppla Update Service implements a preservation recommendation service enabling outsourcing of the selection of appropriate preservation strategies. It provides best practice digital preservation rules for individual collections of Hoppla users.

Preservation Rules of the Update Service. The preservation rules of the Hoppla Update Service are created and administered by a group of preservation experts. A screenshot of the rule administration web site is shown in Figure 2. The rules in the system base on experience in professional settings and intensive testing and evaluation based on test corpus. Unlike in professional settings where preservation rules are created for a specific collection, the rules in the update service are applied to a large number of collections of a specific format. Migrations used in automated migration setting need to provide a high degree of robustness and error tolerance.

Extensive testing and evaluation of the preservation rules is required by preservation experts. For sound testing extensive data corpora of specific formats are required. The objects need to represent the different technical characteristics that a format can have. For example the test objects can vary in size, embedded objects, transparent layer in images, different codecs, etc. The different technical characteristics can have significant effects on the outcome of migrations. Advanced corpora should also contain objects that do not meet the format specification to test the error tolerance of the migration process.

The evaluation and selection of the preservation rules for the web update service is a very challenging task. General preservation requirements and evaluation criteria that apply to all Hoppla users (or even a user group) have to be found.

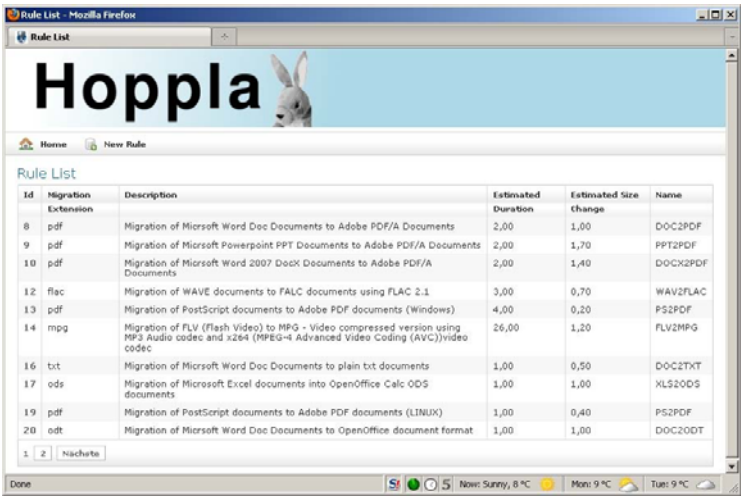

Fig. 2. Web Service Rule Management

Best practices, operating experience of professional archives and de-facto standards (e.g. PDF/A) support the decision making. Through the generalisation not all individual requirements of single users can fully met by the system. Nevertheless the Hoppla system provides best effort preservation rules for specific collections.

As the documentation of the whole process in an archive becomes more important for audit and certification initiatives the web service provides a preservation plan for each preservation rule documenting evaluation criteria, compared alternatives, test objects, potential losses and evaluation results.

A preservation rule of the Hoppla Update Service consist of

- **unique identifier**
- **label & description of the migration strategy**
- **migration tool & applied parameter** including description of tool (licence, operating system)
- **source format and constraints** are specifying in detail the technical characteristics objects covered by the rule need to have
- **target format extension** is required to name the resulting object according to the target format
- **preservation level** specifies the required preservation level (for text, audio, video and images) to apply this strategy
- **estimated duration and size change**, is used to forecast the duration of the migration process and calculate the additional storage usage
- **validity period** of the rule
- **documentation** is a set of documents containing a preservation plan for the preservation rule (evidence of traceability and accountability of the logical preservation in the Hoppla system)

The actual selection of preservation rules for a specific request is based on the collection, user and system profile. In a first implemented version the server selects all rules that are available for the formats specified in the collection profile. After that all rules are selected that conform to the preservation level. Finally, the required tools for the chosen rules are checked whether the tools are installed or available for download from the server, resulting in a list of rules that are send to the client of Hoppla in XML format.

The recommendations provided by the web update service are best effort approaches for automated digital preservation. It can not substitute individual expert planning processes for professional digital preservation endeavours. However, it can provide a practical way for small institutions with limited in-house resources to preserve their digital collections.

3.4 Application of Preservation Strategies

The last step of the workflow is the execution of the recommended preservation strategies on the client side. All objects in the collection that are covered by a preservation rule are identified. The list of migrations is presented to the user. More advanced users can decide which migrations to perform and which to cancel. Furthermore, the duration and additional storage usage of the migrations are estimated based on the figures in the migration rules and the size of the objects in the collection.

The next step is the preparation of the migration services. Portable migration services that are required for the selected migrations are transferred to the Hoppla client. The update service provides a resource based service to download the services, the URLs of the required services are specified in the rules.

The objects in the archive are migrated according to the rules on the client side. Migration is a critical task, as migration services tend to be very error-prone. The large variety of characteristics of a single format that do not always conform to official format specifications (e.g. the use of undocumented features of a format) can cause malfunction of services. Special error handling mechanisms are implemented in the Hoppla software to deal with services that block, abort the migration process or produce unexpected errors. The migrations are executed in a temporary directory on the client system. Successfully migrated objects are added to the existing collection in the data management and metadata about the new objects are collected. In a final step the migrated objects are stored on the storage media of the system.

As privacy is an important aspect of outsourcing it needs to be mentioned that the actual data are not transferred and all actions on the data (migration, identification, metadata extraction) are executed on the client side. Moreover in the next version of Hoppla the level of detail and the containing information of the three profiles (user, collection and system) will be selectable by the user. This should provide the highest degree of transparency for information shared with the web update service.

4 Case Study

A series of two case studies was conducted with a special focus on the logical preservation capacities of Hoppla. In the first case study data of research projects were preserved, the data primarily consists of common office formats. In the second study a business e-mail account was preserved. A common set of rules was applied in both studies with the objective to provide well established and practicable preservation for common office formats (such as text, presentation and spreadsheets). The experiments were executed on workstation PC with a Windows XP operating system.

In our rule set we demonstrate the Hoppla capacity for multiple migration paths for a single format (in our rule set for Microsoft Word documents). The multiple paths should ensure the highest possible support for different future usage scenarios (for example PDF/A for printing and viewing, OpenOffice format for later editing and plain text for text retrieval operations).

Rule set

The following migration rules were applied to both collections:

- **DOC(X)2PDF.** Migration of Microsoft Word documents (objects with .doc and .docx extension) to Adobe PDF/A documents using the Java OpenDocument (JOD) Converter[10] converter. The JOD converter uses an OpenOffice instance[11] to perform the document conversions. On both of our test systems OpenOffice 3.1.1 was installed.
- **DOC(X)2ODT.** Migration of Microsoft Word documents to OpenOffice document format using the JOD converter.
- **DOC(X)2TXT.** Migration of Microsoft Word documents to plain text using the JOD converter.
- **PPT(X)2PDF.** Migration of Microsoft PowerPoint documents to Adobe PDF/A documents using the JOD Converter
- **XLS(X)2ODS.** Migration of Microsoft Excel documents into OpenOffice Calc Files using the JOD Converter
- **FLV2MPG.** Migration of FLV video (Flash Videos) to MPG videos (compressed version using MP3 Audio codec and x264 (MPEG-4 Advanced Video Coding (AVC))video codec) using MEncoder[12]
- **WAV2FLAC.** Migration of WAVE documents to Flac using Flac 1.2[13]
- **PS2PDF.** Migration of PostScript documents to Adobe PDF documents

4.1 Sample Set Office Documents

The sample set contained the data of 5 projects (3 small one and 2 multi-year research projects). The size of the set was 2 GB of data and contained about

[10] `http://artofsolving.com/opensource/jodconverter`
[11] `http://www.openoffice.org`
[12] `http://www.mplayerhq.hu`
[13] `http://flac.sourceforge.net`

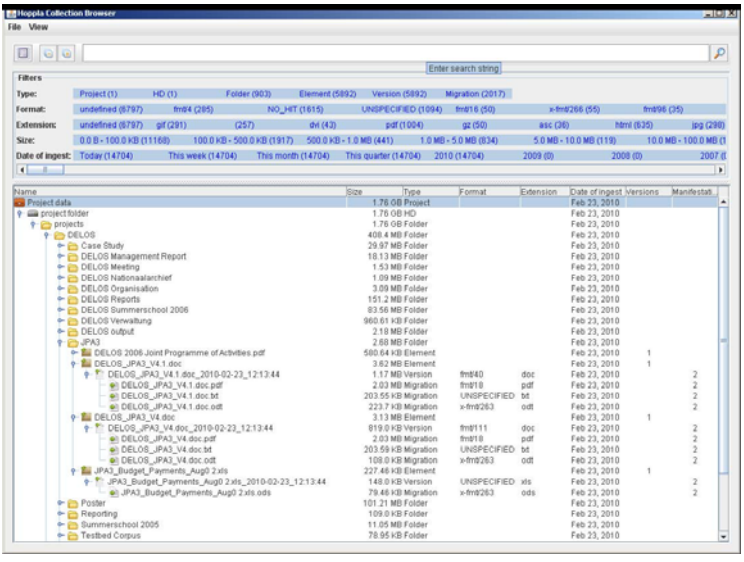

Fig. 3. Collection Browse Projects

6.000 objects in 1000 folders. The collection mainly consists of objects in office formats (e.g. about 524 doc (110MB), 250 xls (160MB), 1000 pdfs (500MB), 500 png (230MB)). Two external hard discs were used as storage media.

Hoppla recommends 2045 migrations and estimated the additional storage need at 660MB. The preservation process (including acquisition, migrations and storage) took about 1,5h. Figure 3 shows the Hoppla collection browser with the resulting collection presenting the multiple migration paths for Microsoft Word documents.

Hoppla successfully migrated 2017 objects, the resulting migrated objects had a size of 380MB. 28 migrations were reported as failure, in particular the migration of PostScript documents to Adobe PDF documents causes ten failures. The implementation of the ps2pdf service on Windows seems to be unstable and required re-evaluation of the rule and the service. The other failures were mainly caused by corrupt documents. The size estimation for the migrated objects was inaccurate and need also further adjustment. Some results in detail, the 524 objects in Microsoft Word format with size of 110MB resulted in 6MB txt documents, 165MB PDF/A documents and 60MB of ODT objects. 248 documents in XLS format (160MB) were migrated into 24MB of ODS format documents. Three documents in flv format with a size of 56MB were migrated to MPG videos according to the migration rules (78MB). The preserved collection on the storage media including the migrated objects and the metadata of the collection in xml format (7MB) had a size of 2,3GB.

4.2 Sample Set E-Mails

The second case study dealt with an e-mail repository containing about 2000 e-mails. About 600 e-mails had attachments, the most common format were Adobe PDF (268), Microsoft Word (145), JPG images (100). The e-mail including the attachments had a size of 271MB, the average size of the attachments was about 300KB. Hoppla stores the e-mail header and content as plain text files. Here, the Hoppla software need to overcome numerousness of different e-mail encodings. The attachments of the e-mails are stored in separate folders. Hoppla identifies the format of the attachment and requested preservation rules from web update service for obsolete formats.

In our case study the Hoppla software performed 493 successful migrations. Eight migrations failed because three documents were corrupt. The migrated object had a size of 73MB. The migration of all objects took about 20 minutes. A manual inspection of some random migration outcomes showed very positive results. The resulting PDF and ODT documents from migrations of Microsoft Word documents were completed and correct. Even the migration to plain text produced satisfactory output. Large parts of the containing text could be extracted.

5 Conclusion

In this paper we described the automation of logical preservation within the Hoppla system. Appropriate preservation strategies for specific collections are recommended by a central web update service. The selection is based on information about the collection, the user and the system which are provided by Hoppla clients. In order to provide migration paths for a wide range of formats Hoppla includes a discovery service to identify potential preservation services on the client side. The recommended migrations are performed on the client side and the migrated objects are managed by the Hoppla system and stored on external media.

Two realistic case studies produced good results and indicated the applicability of the approach. The first set of preservation rules provided migration paths for common office formats. The migration failures were mainly caused by corrupt objects. Only the ps2pdf migration service on windows seems to be unstable and not reliable.

This first version of Hoppla presents the fundamental progress, that will be further refined. The robustness of the migration strategies has to be investigated in more detail and the detection of corrupt objects should be improved. Moreover potential preservation strategies and tools for more formats have to be identified and integrated. A second version of the selection process for migration strategies will allow taking into account more information from the client and thus offering more accurate recommendations.

Acknowledgements

Part of this work was supported by the EU in the 6th FP, IST, through the Planets project, contract 033789 and the Research Studios Austria program of the Federal Ministry of Economy, Family and Youth of the Republic of Austria.

References

1. 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 (2009)
2. Bradley, K.: Digital Preservation: the need for an open source digital archival and preservation system for small to medium sized collections (2008), http://portal.unesco.org/ci/en/files/28067/12323631793BradleyPaper.pdf/BradleyPaper.pdf
3. Ferreira, M., Baptista, A.A., Ramalho, J.C.: An intelligent decision support system for digital preservation. International Journal on Digital Libraries 6(4), 295–304 (2007)
4. Harvard University Library. Global digital format registry (GDFR), http://hul.harvard.edu/gdfr
5. Hunter, J., Choudhury, S.: PANIC - an integrated approach to the preservation of complex digital objects using semantic web services. International Journal on Digital Libraries: Special Issue on Complex DigitalObjects 6(2), 174–183 (2006)
6. Lawrence, G.W., Kehoe, W.R., Reiger, O.Y., Walters, W.H., Anne, K.R.: Risk Management of Digital Information: A File Format Investigation. In: Council on Library and Information Resources (2002)
7. Marketakis, Y., Tzanakis, M., Tzitzikas, Y.: Prescan: towards automating the preservation of digital objects. In: MEDES 2009: Proceedings of the International Conference on Management of Emergent Digital EcoSystems, pp. 404–411. ACM, New York (2009)
8. Oltmans, E., van Diessen, R., van Wijngaarden, H.: Preservation functionality in a digital archive. In: JCDL 2004: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 279–286. ACM Press, New York (2004)
9. Strodl, S., Motlik, F., Stadler, K., Rauber, A.: Personal & SOHO archiving. In: Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL 2008), Pittsburgh PA, USA, pp. 115–123. ACM, New York (2008)
10. The National Archives. Pronom - the technical registry, http://www.nationalarchives.gov.uk/pronom