Migrating Content in WARC Files

Stephan Strodl
Vienna University of Technology
Vienna, Austria
strodl@ifs.tuwien.ac.at

Peter Paul Beran
University of Vienna
Vienna, Austria
peter.beran@univie.ac.at

Andreas Rauber
Vienna University of Technology
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

Heritage institutions all over the world started on harvesting and preserving resources of the World Wide Web for future generations as part of our culture heritage. This task tends to be a non-trivial one because of two complex challenges: (1) crawling the enormous data amount located in the Internet and (2) performing long term preservation strategies on these data. Nowadays a lot of effort is made in the development of Web crawlers and there exist many years' experience with bit storage of large data amounts. However the support for the logical preservation of Internet archives is very limited. The continuous development of technologies that are used in the Web and especially the rapid change in using a tremendous variety of different file formats put the digital assets in the Web archives at risk of becoming inaccessible and unusable in the near future.

This paper presents a workflow to apply digital preservation strategies on the content of WARC archives. The migration of the objects within a WARC archive allows accessing and using the information in the future. The new WARC format that is widely used to store Internet crawl results supports migration of its content. Moreover a set of tools is presented that supports the extraction, migration and injection of objects in WARC files.

Categories and Subject Descriptors

H.3.6 [**Library Automation**]: Large text archives; H.4 [**Information Systems Application**]

General Terms

Design, Documentation, Management, Experimentation

Keywords

Web Archive, WARC, Digital Preservation, Migration

1. INTRODUCTION

Internet archives are crawling, storing and managing enormous large collections of Internet content. These archives contain essential parts of our culture heritages and need to be preserved for future generations.

While crawling and storing these objects is a complex task itself, the Internet archives are facing a new emerging challenge concerning objects in outdated formats. Millions of objects, such as texts, photos, videos, sounds and many more are stored in a remarkable variety of formats within Internet archives. The evolution of these formats danger large amounts of archived data to become inaccessible and not readable in the near future. Whenever the software to render these objects and the documentation describing the internal structure of the formats are not longer available, these objects will become uninterpretable bit streams. No one can guarantee that the required software to render and interpret objects in a Word 95 (document) or Flash (movie) format will be available in 20 or 50 years. Another example is the great variety of video codecs that exist nowadays and are used in the Internet.

Migration can help to provide access on information of harvested objects in the future. Migration, in general, is the conversion of objects from a source format – that is old and at risk of becoming obsolete in the near future – to a target format – a newer format that is easier to handle and access. Continuous migration of objects can ensure access and usage of the information over time. A good example of migration to an easier preservable format is the PDF/A standard [3], which implements a subset of the conventional PDF standard and is especially well-suited for long-time preservation purposes due to its omitting of, for instance, embedded scripts. Current web archiving initiatives are storing their harvested and migrated objects in so called container files. They are hardy aware of the contained object's formats.

The WARC format [4] is a widely-used container format for storing web crawls. It is a revision of the Internet Archive's ARC File Format[1] and was designed for the special requirements of storing web crawls. Moreover the WARC format supports the management of migrated versions of records (so called conversions), which can be directly added to existing WARC archives at a later date. Hereby the use of URIs as global identifiers to address individual objects in WARC

[1] `http://www.archive.org/web/researcher/ArcFileFormat.php`

files allows referring from a migrated object to its original source.

This paper presents a preservation workflow to migrate the content of WARC archives. The aim of the workflow is to provide a toolkit for continuous preservation of WARC files. A set of tools was developed to support the workflow including identification, extraction and verification of objects. These tools allow to perform migration actions and provide simple validation functions for migrated objects. Additional the toolkit supports the extraction of metadata from objects and the injection into WARC files as so called "metadata" records.

The remainder of this paper is organized as follows: Section 2 reveals connections to related initiatives and gives an overview about the work already done in this area. The workflow for content migration of WARC files is presented in Section 3, followed by a description of the software tools in Section 4. A report about a set of experiments is provided in Section 5. An outlook on future developments and conclusions are presented in Section 6.

2. RELATED WORK

In the last years a number of large web archiving initiatives have been initiated. Mainly leaded by national libraries or archives, these organisations started crawling culturally important web content.

The Internet Archive[2] developed the prominent Internet crawler Heritrix[3]. Since 1996 they are using the lossless archive format ARC to store crawl results. The problem was to store large amounts of often small files retrieved from the Internet crawls on conventional file systems. The ARC format provides the storage of simple content blocks sequences representing objects with additional text headers in a self-contained file.

The International Internet Preservation Consortium[4] started a discussion on extending the ARC format for preservation issues. The Danish national web archiving program Netarchive[5] published a report in 2004 [7] that recommends the use of an extended ARC format allowing richer metadata. This report considerably influenced the development of the WARC (Web ARChive) format, superiorly supporting the storage requirements of web crawl results especially for harvesting, accessing and exchanging data. This revision of the ARC format also offers the management of related secondary content, for example assigned metadata or conversions of specific records. This extension is a substantial improvement for the long term preservation capability of WARC files and their content. Hence, the use of these elaborated metadata and conversion records is shown in this paper. Furthermore, in 2009 the WARC file format was accepted as an international standard (ISO 28500:2009).

The Heritrix crawler already supports storing Internet crawl results in this new WARC format. The software approach presented in this paper uses the WARC readers and writers of the Heritrix libraries to handle WARC files. Thus test collection for the experiments presented in Section 5 was crawled by using Heritrix.

The goal of the WARC tool project[6] is to provide documentation, libraries and tools to manage WARC files. An initial set of command line tools is available on the project web site, for example tools to migrate ARC files to WARC files. Further planned developments such as search, validation and extended access tool are still under development and not yet released in a stable version.

Another very prominent tool in the domain of web archiving is NutchWAX[7] (Nutch + Web Archive eXtensions), which is a search engine for web collection archives. It is used within the Internet Archive's Wayback Machine or the WERA (Web ARchive Access) project. In the current version it only supports ARC files.

Long term preservation has become a prominent topic for libraries and archives over the last decade. A lot of effort was spent to define, improve, and evaluate preservation strategies. A good overview of preservation of digital heritage and preservation strategies is provided by the companion document to the UNESCO charter for the preservation of the digital heritage [11].

Research on logical preservation 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 formats, such as e.g. converting a Microsoft Word 97 document into the current Office 2007 format (within format-family migration) or to Adobe PDF/A, a simple ASCII/UNICODE text file, a screenshot image, or another document format. Migration is a modification of the data and thus always incurs the risk of losing essential characteristics of the object [5]. Therefore, a verification of completeness and correctness of the migration activity is required for a preservation system. Characterization services for digital objects that extract information and characteristics can support this verification. Work in the field of characterisation is done, for example, by the Harvard University Library in the JHOVE project[8], the Planets Project with the eXtensible Characterization Language (XCL) [1], and the Global Grid Forum Data Format Description Language Working Group with DFDL [2].

Emulation, the second important preservation strategy aims at providing programs that mimic a certain environment, e.g. the emulation of a certain processor type or emulating the features of a certain operating system. For example running Microsoft Word 1.0 on a Linux operating system by emulating Windows 3.1. Jeff Rothenberg together with CLIR [8] envisions a framework of an ideal preservation surrounding for emulation. Emulation requires sufficient knowledge of the user about the computational environment and

[2] http://www.archive.org
[3] http://crawler.archive.org
[4] http://netpreserve.org
[5] http://netarchive.dk
[6] http://code.google.com/p/warc-tools
[7] http://archive-access.sourceforge.net/projects/nutchwax
[8] http://hul.harvard.edu/jhove

dependencies of components. Emulation of certain software to render data may require preserving the operating system, the application software, and the data. If one of these components fails, the data is lost and the information cannot be accessed or recovered any more. The emulator itself is a piece of software and therefore has also to be preserved over time.

The decision on which preservation strategy (e.g. emulation or migration) to follow is a crucial and complex one and cannot be answered in general. It strongly depends on the individual setting and requirements. In this work we are presenting a workflow for migration of content of WARC archives. The large number of available migration tools and ability to use the migrated data on current systems makes migration an encouraging approach. Moreover, as the WARC format supports the management of migrated objects, it eases the use of migration as preservation strategy for existing Web archives. On the other hand, the additional required storage needs to be seriously considered, especially for Internet Archives.

3. MIGRATION OF WARC

In this section the concept for the migration of WARC records will be explained. Figure 1 illustrates the workflow that consist of four steps:

1. **Preservation Planning**
   Preservation Planning helps to identify and evaluate appropriate preservation strategies. The specific requirements of preserving objects within Web archives need to be fulfilled and the potential loss of characteristics of these objects needs to be documented.

2. **Identification, Extraction and Validation**
   The next step in the workflow is to extract the objects from the WARC container. As the file extension is not a reliable format indicator, other mechanisms are required to determine the actual file format. Also potential validation against the detected format needs to be checked.

3. **Preservation Action**
   Within this step the selected preservation strategies are applied on the validated objects from the WARC records. Wherever possible the resulting objects are checked for completeness and correctness by using characterisation services. The extracted metadata is stored with the objects in the last step of the workflow.

4. **Injection**
   The last step of the workflow covers the generation of a new WARC file. It contains the resulting objects from the applied preservation action with references to the original objects and the extracted metadata.

After the migration and creation of the new WARC files, existing metadata catalogues and indexes of retrieve systems need to be updated. Moreover access tools need to deal with the references between the migrated and original objects. Over the years multiple migration instances of a single object can exist within a WARC archive. Therefore the access engine needs to implement mechanism to provide the preferred version of an object to render.

The following section describes the steps of the workflow in detail. The first task in preservation of WARC records is to identify formats that are at risk of becoming obsolete and require preservation actions. Watch services allow to monitor the technological development of formats, for example the AONSII project [10] monitor format registries. Another promising monitoring strategy for formats can be the analysis of web crawls. The comparison of occurrences of formats between crawls over time can indicate formats that are endangered of becoming obsolete. The monitoring of formats is a continuous process during the lifetime of the Web archive. When a format is identified at risk the preservation workflow is triggered and started with the step 'Preservation Planning'.

3.1 Preservation Planning

The preservation of WARC objects starts with the identification, evaluation and selection of appropriate preservation strategies. Strategies that are used in Web archives need to fulfil specific technical requirements, for example high error tolerance, error handling, batch processing and scalability. Beside the technical aspects of the strategies, the significant properties of the objects need to be specified. Potential preservation strategies need to be tested and evaluated in respect to technical requirements and significant properties. The Planets preservation planning workflow [9] can be used to identify and evaluate potential strategies.

As migration of objects always incurs the risk of losing essential characteristics, preservation planning is used to evaluate potential loss and documents the changes of the objects. The detailed documentation of the preservation strategies and their effects is important regarding traceability and accountability of the archive.

The result of this step is a preservation plan including a description of essential object properties, potential preservation strategies that have been evaluated and the results of the evaluation. The impact on the objects of applying the preservation strategy is also documented in the preservation plan. The plan includes an executable part that specifies the preservation action tool and the specific parameter setting for execution.

3.2 Identification, Extraction and Validation

The next step in the workflow is to identify those objects in the WARC archive that are covered by the preservation plan. First of all, an index of all available records in the WARC archive needs to be constructed, if it does not already exist.

All content blocks i.e. the files of WARC records with the format extension specified in the preservation plan are extracted from the WARC archive. As the file extension is not a reliable format indicator, the objects need to be identified by using format identification services. The software tool supporting this workflow (described in Section 4) uses the format identification tool DROID[9] from the National Archive. It allows the identification of the precise format

[9] `http://droid.sourceforge.net`

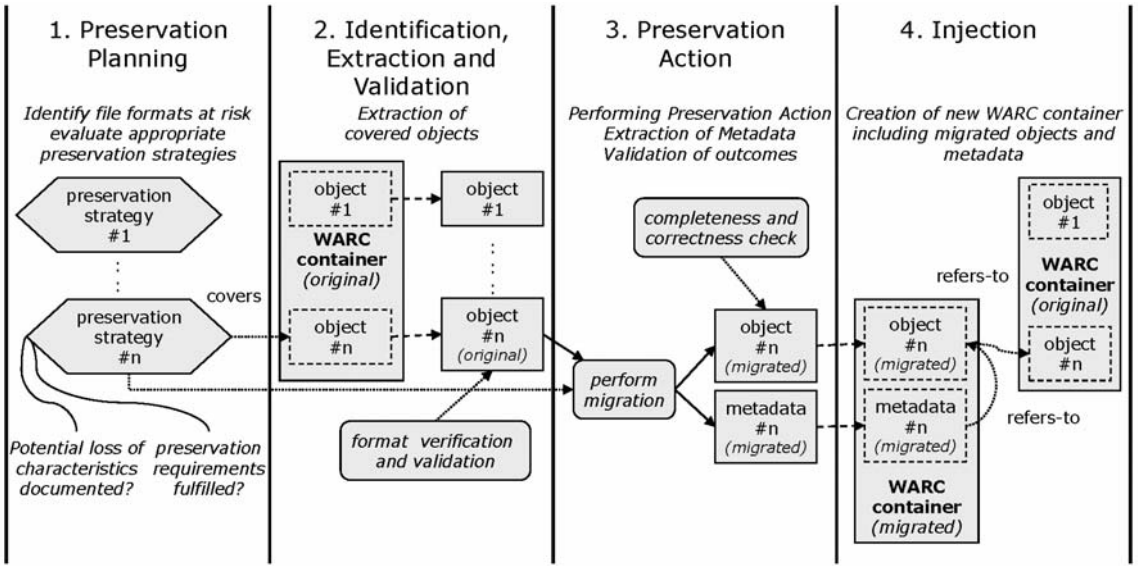

Figure 1: Workflow for migration of WARC records

(e.g. format including the version number) of objects. Where appropriate services are available, the extracted objects are also validated against the identified format specification.

The results of this step are the extracted, identified and verified objects in the format that is defined in the preservation plan as the input format.

3.3 Preservation Action

The preservation action step is responsible for migrating the verified objects using the tool and parameter setting defined in the preservation plan. Where possible and appropriate tools are available the result of the migration should be verified for completeness and correctness.

The tool support for validation of the correctness of migrated objects is limited, especially for migration across format families (for example Word document to TIFF images). Most of the characterisation tools only verify the structure of the objects and check whether the output is well-formatted according to a format specification. Hence, this only allows a very minimalistic quality check of the migration outcome, missing the validation of completeness and correctness of the contained information.

Afterwards migration characterisation services are used to extract additional metadata form the objects. Metadata is vital information for search and retrieval functionality as well as for continuous preservation activities. An example of such a characterisation service is the JHOVE project. The result of this step is a list of migrated and (where possible) verified objects with additional metadata.

3.4 Injection

The last step of the workflow adds the migrated objects to the WARC archive. Therefore a new WARC file is created that stores all migrated objects as a WARC record of WARC-Type `conversion`. The WARC field `WARC-Refers-To` points to the record containing the original object by using its record id. Listing 1 shows the record header of a migrated object. The new target URI is the URI of the original record with a new file extension. For example the Word document at
`http://www.tu-sofia.bg/Bul/norm-dok/prav-UEP.doc`
was migrated to a PDF document and the new URI
`http://www.tu-sofia.bg/Bul/norm-dok/prav-UEP.pdf`
was assigned to the migrated record.

Listing 1: Conversion Record of migrated object in a WARC File

```
WARC/0.18
WARC-Type: conversion
WARC-Target-URI: http://www.tu-sofia.bg/Bul/
norm-dok/prav-UEP.pdf
WARC-Date:
2009-07-22T11:08:24Z
WARC-Refers-To: <urn:uuid:1b5d742f-2f6c-4f03-
8a79-3dd9551b9570>
WARC-Record-ID: <urn:uuid:cf7e6b9a-4f26-447b-
9a2a-25c0cc5e419e>
Content-Type: txt/pdf
Content-Length: 64799
...
```

A common practice for long term archives is to store the metadata together with the migrated objects. The metadata are stored within the new WARC file as a WARC record of WARC-Type `metadata`. Listing 2 shows the record header of the metadata record, which contains the URI of the migrated object and points via the `WARC-Refers-To` field at the migrated object. As the content block of a metadata record has no specified structure, the record varies between different implementations and tools. In our implementation

we used a XML structure for the metadata content block, which should ease the use of the data in a machine-driven automatic way. The example in Listing 2 shows the first entry in the content block which is the output from DROID, followed by the output from JHOVE.

Listing 2: Metadata Record of migrated object in a WARC File

```
WARC/0.18
WARC-Type: metadata
WARC-Target-URI: http://www.tu-sofia.bg/Bul/
norm-dok/prav-UEP.pdf
WARC-Date: 2009-07-22T11:08:24Z
WARC-Refers-To: urn:uuid:cf7e6b9a-4f26-447b-
9a2a-25c0cc5e419e
WARC-Record-ID: <urn:uuid:21f0b396-c3f4-4e84-
a2a6-f319b6cc6182>
Content-Type: text/xml
Content-Length: 14462

<metadata>
 <output tool="DROID_3.0.0_Signature_File_V16">
    fmt/18
 </output>
 <output tool="JHOVE_1.1">
   <?xml version="1.0" encoding="UTF-8"?>
...
```

The result of the workflow is a set of new WARC files for each defined and executed preservation strategy. These files include the migrated objects as "conversion" records each of them referring to its original record in the origin WARC file. Moreover, all metadata captured during the workflow including format identification are stored in the new WARC file as metadata records.

4. SOFTWARE

A set of tools was developed to support the workflow described in Section 3. The toolkit provides software support for the steps 2 to 4 of the workflow. The software is developed in Java and allows to handle WARC files including indexing, extraction and creation by using the tools and libraries provided by the Heritrix project, especially the `WARCReader`/`WARCWriter` classes. Before using the toolkit the first step 'preservation planning' of the toolkit needs to be done. Hereby the planning software Plato[10] can help to create an appropriate preservation plan. The software toolkit consists of the following components similar to the workflow.

Extraction Tool: If necessary the extraction tool creates an index of the records in the WARC archive, otherwise it uses the existing index. The content blocks are extracted of all records with the format extensions defined in the preservation action setting (described below). The extracted files are currently stored in a temporary directory. In a more sophisticated version of our tool this needs to be optimized via a processing pipeline. Our approach is using DROID to determine the precise format of the objects (e.g. PDF version 1.4). Additional for those formats that are supported by JHOVE, the objects can be validated against the format specification.

Migration Tool: The migration tool performs the preservation actions defined in the preservation action settings on the extracted objects. In the current version of the software, the preservation actions are performed using system commands. The system command executes the tool that needs to be installed on the host system. After the migration, the characterisation services DROID and JHOVE are used to identify and verify the format of migration outcome and obtain additional metadata.

Injection Tool: For each executed migration plan a new WARC file is created containing all migrated objects referring to the original object and the extracted metadata record in XML format. Examples of record headers for migrated objects and metadata are shown in Section 3.4.

Preservation Action setting

All settings that are used by the toolkit are defined in the preservation action setting. The setting is part of a preservation plan resulting from preservation planning activity. The preservation action setting are used by the toolkit consists of the following elements:

- **Name** as a short description of the preservation plan.
- **Description** is added to the resulting WARC file (as a content description in the WARC info header) and should ensure that the content of the WARC file can be understood in the future. It usually includes the outcome of the preservation planning process, the description of the tool and the used parameter setting.
- **Input Extension** defines the format extension of the input objects. It is used to extract the records from the WARC archive.
- **Output Extension** specifies the format extension of the migrated object.
- **DROID Input Id** defines the DROID id for the input objects and is used for exact verification of the objects. As DROID supports only a limited number of formats, the DROID attribute is optional. Otherwise only the file extension is used to identify the input objects.
- **System Command** is executed on the system to perform the migration and specifies the tool to use and the command arguments.
- **DROID Output Id** can provide the DROID id for the migrated object that can be used for a minimalistic verification of the migration outcome.
- **Output Mime Type** defines the mime type of the migrated objects and is used in the WARC record header.

[10] `http://www.ifs.tuwien.ac.at/dp/plato`

5. EXPERIMENTS

An initial set of experiments was performed to demonstrate the feasibility and provide a first estimation of resources consumption. The WARC archives were crawled with Heritrix. The test data for both settings were crawled from a Web site[11] presenting the work in digital preservation of the Vienna University of Technology and parts of the Web site from the Technical University of Sofia[12]. The resulting WARC file contained 2485 documents and had a size of 197 MB.

5.1 Image Migration

The first experiment setting tests the migration of images by using the convert tool of ImageMagick[13].

Therefore, two example preservation action settings were defined. The first setting represented a migration approach to convert GIF-images to PNG-images. The second approach migrated JPG-images to PNG-images. The crawl contained 165 GIF-images with a size of 1.1 MB and 480 JPG-images with a size of 15.1 MB. The resulting changes in size are presented in Table 1.

Input		Output		WARC
Format	Size	Format	Size	Size
GIF	1.1 MB	PNG	1.1 MB	1.2 MB
JPG	15.1 MB	PNG	47.6 MB	70.1 MB

Table 1: Results of the Image Migration

The migration of GIF-images to PNG-images resulted in a negligible change in size and took about one minute. The tool reported an error caused by a GIF-image that had a transparent layer. It could not correctly be migrated by the convert tool.

Four objects extracted from the data set with the JPG-file extension were no JPG-images and had been detected by DROID. The migration from actual 476 JPG-images to PNG-images resulted in a change in size from 15.1 MB to 47.6 MB, the resulting WARC file had a size of 70.1 MB. The execution of the workflow took about four minutes.

All resulting objects could be verified by DROID as PNG-images. Furthermore, no corrupt objects were found through a manual inspection of all migrated objects.

5.2 Word Migration

The second setting was the migration of Microsoft Word objects to PDF objects using the Java OpenDocument Converter[14]. The tool uses an instance of OpenOffice for the migration. The test corpus contained 193 crawled Microsoft Word objects with a total size of 8.6 MB. The word objects were primarily small forms with only one or two pages.

The migration of the objects includes the verification of the resulting format by using DROID and the extraction of metadata by using JHOVE. The experiment took about five minutes. The resulting PDF objects had a size of approximately 10 MB and the resulting WARC files were 13.2 MB large (shown in Figure 2). JHOVE extracts very detailed metadata information from PDF documents that results in large metadata WARC records. A manual random quality control showed correct migration results. DROID and JHOVE verified all migrated objects as well formatted PDF documents in Version 1.4.

Input		Output		WARC
Format	Size	Format	Size	Size
DOC	8.6 MB	PDF	10.0 MB	13.2 MB

Table 2: Results of the Word Migration

A second execution of the migration workflow skipping the characterisation services took only a few seconds. In this migration setting the extraction of metadata using JHOVE is the most time-consuming task.

As depicted above, the experiments provided a first positive proof for the applicability of the tool set. It revealed that the migration workflow is a relative time consuming process even for a small set of test objects. In particular the characterisation services took a lot of time to analyse objects and extract their properties.

Experiments with larger data sets are required for detailed analysis of the scalability of the tools. Additional required storage capacities for the migrated object as well as for the metadata need to be serious considered for migration activities of large amounts of data.

6. SUMMARY AND OUTLOOK

In this paper we discussed the logical preservation of WARC archive content. Long term preservation of Web archives is an important issue to enable future access and usage of crawled information.

A workflow for migrating the content of WARC archives is presented in this paper. It covers preservation planning, extraction and migration of objects and the injection of new WARC files. Moreover, the workflow assists the extraction of metadata from the migrated objects and injects the obtained data to existing WARC archive.

Furthermore, a set of tools supporting the workflow is presented in this paper. The toolkit demonstrate the use of the specific WARC data types for metadata and migrations. A first experimental setting allows an feasibility evaluation of the software, but further experiments with larger data sets are required to evaluate the scalability of this approach.

The support of access engines for migrated records and extracted metadata needs to be further analysed. Another research aspect is the migration of complex objects. Here, the effects of migrating sub-elements of complex objects and re-rendering of complex objects with migrated sub-elements needs some further research effort.

[11] http://www.ifs.tuwien.ac.at/dp
[12] http://www.tu-sofia.bg
[13] http://www.imagemagick.org
[14] http://www.artofsolving.com/opensource/jodconverter

7. REFERENCES

[1] C. Becker, A. Rauber, V. Heydegger, J. Schnasse, and M. Thaller. A generic xml language for characterising objects to support digital preservation. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing*, New York, NY, USA, 2008. ACM.

[2] M. Beckerle and M. Westhead. GGF DFDL Primer. Technical report, Global Grid Forum Data Format Description Language Working Group, 2004.

[3] ISO. *Document management - Electronic document file format forlong-term preservation - Part 1: Use of PDF 1.4 (PDF/A) ISO/CD 19005-1*, 2004.

[4] ISO. *Information and documentation – WARC file format (ISO 28500:2009)*, 2009.

[5] G. W. Lawrence, W. R. Kehoe, O. Y. Reiger, W. H. Walters, and K. R. Anne. *Risk Management of Digital Information: A File Format Investigation*. Council on Library and Information Resources, 2002.

[6] G. W. Lawrence, W. R. Kehoe, O. Y. Rieger, W. H.Walters, and A. R. Kenney. Risk management of digital information: A file format investigation, June 2000.

[7] Netarkivet. Archival data format requirements. `http://netarchive.dk/publikationer/Archival_format_requirements-2004.pdf`, July 2004.

[8] J. Rothenberg. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Council on Library & Information Resources, 1999.

[9] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)*, pages 29–38, New York, NY, USA, 2007. ACM.

[10] The National Library of Australia. Automatic obsolescence notification system (AONS). `http://pilot.apsr.edu.au/wiki/index.php/AONS_II`.

[11] UNESCO. *Guidelines for the preservation of digital heritage*. UNESCO, Information Society Division, March 2003. `unesdoc.unesco.org/images/0013/001300/130071e.pdf`.