# Package Formats for Preserved Digital Material

Eld Zierau

The Royal Library of Denmark
Søren Kierkegaards Plads 1
1219 København K
ph. +45 33 47 46 90

elzi@kb.dk

## ABSTRACT

This paper presents an investigation of the best suitable package formats for long term digital preservation. The choice of a package format for preservation is crucial for future access, thus a thorough analysis of choice is important.

The investigation presented here covers setting up requirements for package formats used for long term preserved digital material, and using these requirements as the basis for analysing a range of package formats.

The result of the concrete investigation is that the WARC format is the package format best suited for the listed requirements. Fulfilling the listed requirements will ensure mitigating a number of risks of information loss. Thus WARC is the best choice for a package format in cases where these same risks are judged most important. Similar analysis will need to be carried out in cases where the requirements differ from the ones described here, e.g. if there are specific forensic or direct access to files.

## Categories and Subject Descriptors

E.2 Data Storage Representations: Linked representations, Object representation

E.5 Files: Backup/recovery, Optimization, Organization/structure

H.3.7 Digital Libraries: Collection, Standards, Systems issues

I.7.1 Document and Text Editing: Document management, Version control

I.7.2 Document Preparation: *Format and notation*, *Standards*

## General Terms

Management, Documentation, Design, Standardisation.

## Keywords

Package formats, Digital Preservation, Bit preservation.

## 1. INTRODUCTION

This paper presents an investigation of different possible package formats that can be used for packaging digital material for long term preservation. The investigation has resulted in suggesting the WARC format as the package format to be used for bit preserved digital material at The Royal Library of Denmark [2].

The selection of a package format for digital material is crucial for how to facilitate long-term accessibility. The selected package format is used to package files that must be sent to bit preservation, which must ensure that the bit-streams remain intact and readable [11,25]. That means the package format will constitute the frame of the digital material, and thus be the basis for general recovery of data and future data access as well as functional preservation actions of the original bits, where functional preservation ensures that the bits remain understandable and usable according to the purpose of preservation [25]. A package formats is presumed needed, because files must be applied a minimum of metadata in terms of an identifier as described later.

The topic of long term preservation package formats has partly been treated in a recent paper: “Digital forensics formats: seeking a digital preservation storage format for web archiving” [10]. As the paper states: “There has been little consensus on best practices for selecting storage container”. The paper presents an overview of archiving formats for digital forensics that can satisfy the requirement of tracing originality. This present paper on the other hand will not focus on requirements for forensics, but instead will focus on requirements for long term preservation in general.

The goal of the investigation was to find as few suitable package formats for packaging for as many types of different materials as possible. The reason for this goal is that each package format will require resources in form of skills and documentation in order to maintain accessibility to the material. Thus in order to minimize costs and in order to minimize the risk of losing skills for a specific format, the number of formats must be kept as low as possible.

Diverse types of digital material can for instance be found for libraries. Libraries usually have many types of different digital materials that are candidates for long term preservation. For instance substitution copies of analogue materials [9]; harvested web material [2]; emails from authors and forensic images of e.g. author’s hard discs [10]. The digital material can consist of different files with different file formats and metadata, and the material can be composed digital objects (called representations as in PREMIS terminology [17]) with various metadata.

This paper will argue for a set of requirements that should be considered in choice of a package format used in long term preservation of diverse types of digital material. Such

requirements will depend on the purpose of the preservation, the nature of the material to be preserved and individual prioritization of risk that must be mitigated by the way the material is preserved. Thus, the given requirements are arguable requirements to be considered, while the weight of meeting them can differ.

The next section will provide the general requirements for a package format used for long term preservation of digital material. The following section 'Alternative package format choices' describes a range of packaging formats and analyses how they meet the different requirements for a package format.

# 2. FORMAT REQUIREMENTS

The format requirements described here are the requirements for formats used for archive packages under long term preservation. The following contains descriptions and arguments for a number of such requirements. These requirements are either related to the actual packaging and storage, to preservation aspects, or to identification of contents of packages.

## 2.1 Package and storage related requirements

The following requirements are requirements related to packaging and storing. These are selected requirements which cover the most often referred requirement about independence, as well as requirements related to flexibility concerning exploitation of storage resources. More detailed requirements are left out in order to give a comprehensive presentation (additional requirements can e.g. be found in [2,10]).

***Requirement 1: Independence of storage platform***

For long term bit preservation, data will in most cases be stored on different media using different operating systems. This is, for instance, the case for one material in order to ensure independence between copies of data in a bit repository, which takes care of holding and preserving bits [25]. In the long term this is likely to be the case at some stage as a consequence of changes in storage technology. Thus a basic requirement for a package format used in long term bit preservation is: *The Package format is independent of storage platform* [2], which has been formulated in many ways as a requirement for sustainable file formats in general [2,10,12,13,14,22].

***Requirement 2: Package format allows flexible packaging***

A requirement related to how well the format can support optimization of storage use is: *Package format allows flexible packaging.* This can relate to economical or performance related issues concerning the best way to package, making different sizes of packages. There can be benefits in having large packages according to how the storage works. On the other hand there can be accessibility issues which can mean that smaller packages are preferred. Reasons to keep to small packages can be technology changes as well as challenges in having different parts of the packages with e.g. different confidentiality levels. Anyhow, flexibility will mean that package sizes can be optimized according to chosen policies[1].

[1] Discussion on this subject is documented in mail correspondence with Kevin Ashley on the JISC Digital Preservation mailing list. Please refer to https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1105&L=digital-preservation&F=&S=&P=7686

***Requirement 3: Allow update records***

A requirement related to the ability to minimize needed storage volume is to require that the: *Package format allows update records*. Since data packages for long term preservation are static, they cannot be changed after bit preservation has started. Therefore the only alternative to update packages is to make a full new representation and bit preserve this representation as well. However, in many cases this can be expensive, for instance in the case where a large TIF file has a single letter change in the TIF file header. However, the opportunity of having update records must be carefully considered in terms of the complexity it can add to the long term interpretation of the data.

## 2.2 Preservation related requirements

Preservation related requirements for package formats cover aspects of ensuring that the packages are readable and understandable in the future. These have many similarities to general requirements for preservation file formats [12,13,14,18,22]. Common to such requirements is that they are related to mitigating risk such as losing information in the digital material or losing ability to interpret the information [20,24].

The following requirements are deduced from an analysis where risks and requirements are considered for digital material that will have a large variety and will have to be long term preserved. These requirements are based on the above mentioned literature and further details can be found there as well.

***Requirement 4: Must be Standardised format***

The first requirement is that it: *Must be a standardised format.* This covers the degree to which the format has gone through a rigorous formal standardisation process [12,13,14,18,22]. This relates to the future ability of thorough and accepted documentation for the format which will mitigate risk of losing means to understand the format.

***Requirement 5: Must be open***

A related requirement is that a format: *Must be open* [2,14,18,22]. This requirement relates to risks of losing the ability of future interpretation of the format. If the format is not open, there may arise legal and economical issues concerning tools to interpret the contents of the format. Furthermore, there may be a risk that documentation of the format is unavailable after e.g. copyrights of the format have expired.

***Requirement 6: Must be easy to understand***

Another related requirement is that the format: *Must be easy to understand.* This requirement is usually referred to in connection with transparency [2,12,13] and complexity [6]. The requirement relates to the future ability to understand the package format, and to mitigate the risk of introducing errors or later difficulties in interpreting the contents of packages. This risk is high if the format is too complicated.

***Requirement 7: Must be widely used in bit repositories***

There is a requirement stating that the format: *Must be widely used in bit repositories*. This covers ubiquity in terms of the extent to which the format has been adopted. In particular in this paper *widely used in bit repositories* means the extent of adoption by national libraries, archives, and other memory institutions internationally [12,13,14,18,22].

***Requirement 8: Must be supported by existing tools***

A related requirement is that the format: *Must be supported by existing tools.* This also concerns the trust in quality and future existence of the format, which again will mitigate the risk of losing ways to understand the format in the future. Furthermore it concerns the ubiquity aspect in terms of how widespread the format can become [14,18].

***Requirement 9: Must be able to include digital files unchanged***

The final preservation format related requirement is that the format: *Must able to include digital files unchanged.* This requirement addresses mitigation of the risk of losing information as a result of changes made to files in the packaging process. Such changes could for instance occur in connection with compression (partly discussed in [12,22]). Or in cases where the package format is XML based, and conversions are needed in order to include files in XML structures due to the fact that XML is tag based, and end tag can be part of the files.

## 2.3 Identification related requirements

The last requirement covered in this paper is a requirement related to the ability to identify contents of packages, which is the basic metadata of any digital piece of information.

***Requirement 10: Must facilitate identifiers for digital files***

The requirement that a package format: *Must facilitate identifiers for digital files.* This requirement is related to more general requirement of flexibility of embedding metadata [10]. It does however deserve special attention and explanation, since it is crucial for future reference of files which are part of digital material.

In general we have three different types of data which must be recorded in packages. The three different types of data[2] are:

- *Digital files* of any file format will need to be addressed in different contexts, such as metadata for the file or relations to the files as part of a digital object. Therefore the digital files must be identifiable. This is done by assigning an identifier to each file.
- *Metadata to digital files* as metadata about the files separated from the actual files. This metadata will as a minimum consist of the identifiers for the digital files.
- *Metadata for a representation*. All information for contexts and metadata can be put into e.g. a METS[3] structure with references to the involved files and metadata.

These types correspond to the object types 'file' and 'representation' in the PREMIS metadata standard, where a representation can be purely representation of file metadata.

Different metadata schemes facilitate definition of identifiers for the metadata, thus it is no problem to make schemes of how to represent identifiers for and within the metadata. However, definition and attachment of usable identifiers for digital files is a challenge, since the digital file itself may not carry the information of the identifier of the file.

One solution to meet this challenge could be to simply place the files as bit chunks with the identifier to the bit repository, and leave it to the bit repository to make the connection between the file and the identifier. However the information that the file has been assigned the specific identifier is also crucial for long term preservation. If we leave it 100% up to the bit preservation solution to preserve the link between files and identifiers, we will risk that we cannot recreate the data in case this index is lost. Furthermore, if the identifier is only expressed as an identifier in a bit repository, we eliminate any optimisation of packaging more files or files and metadata in the same packages for a bit repository. Therefore the best way to ensure the relation is to put the identifier with the file.

There are different ways to assign information of an identifier with a file:

- *Naming files with the identifier*
  Using identifiers in file names is generally not considered a good solution, for a number of reasons:

  Firstly, because there can be restrictions to how files are named which can conflict with the general scheme to name persistent identifiers.

  Secondly, because a file name is not part of the file itself, it is information of the file system. Furthermore, the file name can only be unique in connection with a file path anyway, and a file path will include an assumption on how files are placed which is likely to change in a time frame of 50 years. This again can give challenges to update of reference and resolver schemes.

  Thirdly, file names may not make sense in the future, and in a bit preservation context with different copies on different media as e.g. microfilms, file names may not exist or may be different for different copies in a bit preservation system.
- *Put identifier into files as inherited metadata*
  Insertion of an identifier into files would have to be done before the files are sent to bit preservation. This could work for some cases, but cannot be used in all cases. First of all because not all file formats allow inherited metadata. Secondly, because there may be requirements to leave the file untouched (e.g. a forensic disc image). In general it would also require knowledge of how to extract the identifier from all bit preserved file formats, which in practice would not be possible for collections with all types of digital material.
- *Wrap files and identifier in a package format*
  Wrapping an identifier with the file in a package will set requirements for the abilities of the package format, since this is not a trivial feature that applies for all package formats.

This requirement of facilitating identifiers for digital files is therefore based on the assumption that we want to mitigate the risk of losing identifier information because of environment or file format dependencies.

# 3. FORMAT CHOICES

This section describes a range of different package formats that could be candidates for a general package format for a wide range of digital material, as is usually the case for libraries. This section will furthermore describe how well the formats fulfil the different requirements listed in the previous section.

[2] Except from the metadata part, this corresponds to different types of PREMIS objects [16]

[3] Metadata Encoding & Transmission Standard (METS) http://www.loc.gov/standards/mets/

## 3.1 Considered package formats

The following considered package formats are chosen based on knowledge of package formats used in other libraries and archives repositories[4], formats described in the paper "Digital forensics formats: seeking a digital preservation storage format for web archiving" [10], and generally known package formats such as ZIP and RAR. The list of formats does not constitute an exhaustive list of formats. For instance the Archive eXchange Format (AXF)[5] is excluded since "… it is a very new development, with a lack of access to detailed documentation and source code, making it difficult to assess" [10]. Also formats for very specific purposes like the optical media disk imaging format iso image are excluded [8], and the format gzip[6] which is a compression format and thus cannot fulfil the requirement of unchanged files. In order to narrow the list, there are also formats that are described together with other formats, which for instance is the case for XFDU which is mentioned under METS.

### *3.1.1 AFF*

Advanced Forensic File Forensic disk image formats such as AFF[7] and AFF4[8] are formats specifically designed for to contain metadata for forensics. These formats have the benefit of providing settings to control the quality, speed, and size of output data. One disadvantage of AFF is that it assumes that the image is from a disk as opposed to a collection of files or folders [10].

Take for example the AFF4 format, an open format which is proposed to be adopted as a standard evidence management platform [3]. The AFF4 is a position based format with the ability to insert specific forensic metadata. However it does not support means of update records.

### *3.1.2 ARC*

The ARC format is a position based format originally designed for web archiving packages. It is based on record definitions identified by name tags and byte length. It requires that the first record in a package is a header record, a 'filedesc' record, with information that is only used in the context of web archives and thus can add confusion and take up space for packages that are not web archive specific[9] [11].

The ARC format has a fixed set of record definitions, i.e. it does not include the possibility to define separate update records. The ARC format is not described in a standard and it is not very widely used for other archives than web archives. Furthermore, there is a tendency that web archives using ARC are moving to use WARC instead [23].

### *3.1.3 BagIt*

The BagIt[10] format is intended for quick packing and unpacking into folders. It was originally design for exchange of information, i.e. BagIt is not directly designed for packaging to archives. The BagIt format only provides a way to specify certain metadata to a package, whereas the package itself must be specified to be a package in e.g. TAR or ZIP formats.

The BagIt format provides a structure for how files can be packed in e.g. a TAR or a ZIP file. It allows for specification of one external identifier, but otherwise it does not offer other ways to address the files in the bag aside from their file names.

The BagIt format is used both as exchange format but also as a package format for data in a repository[11]. The BagIt format is not formally standardised. The BagIt format cannot be extended with support of update records.

### *3.1.4 METS*

The Metadata Encoding and Transmission Standard (METS) specifies an XML based format which originally was designed for transmission of information, but is today widely used as a container format for metadata to digital material[12] [22].

The METS format could in theory be used as a package format, although there are challenges regarding inclusion of digital files in a METS structure. The challenge is due to the fact that METS is an XML based format and in practice XML is not suited for inclusion of digital files, since objects are defined via start and ending tags. Thus the file will need to be transformed in order to avoid ambiguity in case the file itself includes bit sequences that can be interpreted as an end tag. This is probably the reason why METS is often used as metadata format but rarely used as the actual package format (examples of METS packed in WARC or BagIt can be found in [5] and [4]).

The METS format is very flexible and can include a range of other XML based metadata formats. It may therefore be possible to exploit this flexibility to include specification of update records. The METS format is a widely used standard hosted at the Library of Congress[13]. However, the standardisation is related to METS as a metadata standard rather than a package format standard.

Another similar format is the XFDU format [1], also an XML based metadata format. The XFDU format therefore has the same challenges as METS also being based on XML.

---

[4] Partly based on the previously mentioned mail on the JISC Digital Preservation mailing list

[5] See http://www.openaxf.org/ for description of AXF

[6] The gzip fomat is defined in "GZIP file format specification version 4.3", http://www.ietf.org/rfc/rfc1952.txt

[7] See description of Advanced Forensics Format (AFF) on http://www.forensicswiki.org/wiki/AFF

[8] See description of Advanced Forensics Framework 4 (AFF4) on http://www.forensicswiki.org/wiki/AFF4

[9] See "Arc File Format, Version 1.0", http://www.archive.org/web/researcher/ArcFileFormat.php

[10] The BagIt fomat is defined in "The BagIt File Packaging Format (V0.97)", http://tools.ietf.org/html/draft-kunze-bagit-06

[11] See e.g. http://www.dcc.ac.uk/resources/external/bagit-library

[12] See e.g. "METS Implementation Registry", http://www.loc.gov/standards/mets/mets-registry.html

[13] See http://www.loc.gov/standards/mets/

### 3.1.5 RAR

RAR stands for Roshal ARchive. It is a proprietary archive file format that includes data compression[14]. The RAR format is not an open format and it is not formally standardised.

RAR files may be created only with the commercial software WinRAR, RAR, and software that has been granted permission.

The RAR format is mainly focused on technical issues related to the actual storage of packages in compressed form. It does not provide means to specify external identifiers and there are no possibilities of making extensions with update records.

### 3.1.6 TAR

The TAR format[15] provides a way to package file folders and their contents. The TAR format is file oriented, but also byte oriented. The TAR format has no centralized location for the information about the contents of the file, i.e. it is not easy to make relations between identifiers and files. The best way to assign identifiers to TAR elements is to use the BagIt format which opens more possibilities to specification of different data.

The TAR format is a standardised (POSIX.1-2001) format which is widely used for archiving of tapes in general, and there are different tools available for the format. The TAR format does not support the notion of update records.

### 3.1.7 WARC

The WARC format is a position based format focused on web archiving, but has a general design which can also be used for other purposes, leaving out web specific information [7].

The WARC format consists of different record types, where a record e.g. can contain a file as well as record information as for instance the identifier for the record/file. Thus WARC provides an easy way to assign an identifier to a file.

The WARC format has recently been ISO standardised [7], but is not used very widely yet and there are few tools available. WARC has recently been used for other material than web material in the German Kopal project [21].

As for the ARC format, the WARC format also has header information, but in this case it can consist of information that is relevant for a bit repository, including the identifier for the package itself.

There have been initiatives to develop tools for WARC in different contexts: at the University of Maryland[16], in an IIPC project[17], and at Internet Archive[18]. However, these tools are still not mature enough to consider as proper production tools [15].

The standard includes the possibility to define your own record type [7], which enables us to specify updates as basis for update mechanisms.

### 3.1.8 ZIP

The ZIP file format[19] is a file format, which is used for data compression and as an archive format, which also allows for uncompressed packaging. A ZIP file can contain file folders and files. For each entry there are defined a number of fields like file name, compression algorithm etc. The format also allows specification of additional fields, e.g. the identifier for a file.

The ZIP format was originally published as an open format [16]. Although ZIP is widely used in general and proposed to be standardised, it has never been formally standardised[20]. Furthermore it should be noted that although ZIP is widely used in general, it is not as common to see ZIP used as package format in archives and libraries.

There are different implementations and interpretations of the ZIP format [10]. Exploiting the ability to define an identifier in an extra field would also require specifically design zip tools to make this information extractable.

The ZIP format does not have any direct mechanism enabling introduction of update records.

There are different software components deployment formats building on ZIP, e.g. the Web application ARchive (WAR)[21] file format, and the Java Archive (JAR)[22] file format. As these formats are designed for software deployment rather than for archiving, these formats do not provide extra means for archiving than the ZIP format.

## 3.2 How the formats meet requirements

An overview of how the presented package formats meet the requirements for the package format used in long term preservation is provided in table 1. The table provides approximate ranking of how well the formats meet the requirements. These rankings are expressed by the five ranking values (illustrated by colours in order to give a better overview):

*Yes* if the requirement is considered to be sufficiently met

*Almost* if the requirement almost can be considered to be sufficiently met, but not completely

[14] See "RARLAB" for description of the RAR format http://www.rarlab.com/

[15] Description of the tar file format can be found on http://www.gnu.org/software/tar/

[16] See "An Approach to Digital Archiving and Preservation Technology – WarcManager", https://wiki.umiacs.umd.edu/adapt/index.php/WarcManager

[17] See "Open Source WARC Tools - Functional Requirements Specification", http://warc-tools.googlecode.com/files/warc_tools_frs.pdf

[18] See "Release Notes - Heritrix 3.1.0-RC1", https://webarchive.jira.com/wiki/display/Heritrix/Release+Notes+-+Heritrix+3.1.0-RC1, retrieved October 2011

[19] See "ZIP File Format Specification" http://www.pkware.com/documents/casestudies/appnote.txt

[20] See http://www.itscj.ipsj.or.jp/sc34/open/1414.pdf which proposes standardisation.

[21] See e.g. "Web Application Archives" for description of the Web ARchive (WAR) file format (Sun), http://java.sun.com/j2ee/tutorial/1_3-fcs/doc/WCC3.html

[22] See e.g. "JAR File Specification" http://docs.oracle.com/javase/6/docs/technotes/guides/jar/jar.html

*So-So* if the requirement is considered to be met to some extent, but thorough evaluation of deficiencies is required

*Little* if the requirement is only considered to be sufficiently met to a minor degree

*No* if the requirement is not considered to be met at all

The ranking is only approximate values, since e.g. definition and evidence of whether formats are widely used are only based on knowledge of a small set of larger institutions. It should also be noted that there is an emphasis of use of the formats as package formats in preservation, thus the METS format is rated to be 'so-so' *widely used in bit repositories*, since it is widely used as a metadata format, but not as a package format. Likewise the ZIP is ranked 'so-so', since the requirement concerns the widespread use of ZIP with bit repositories for long term preservation in larger preservation institutions. Another example of approximation is that the BagIt format cannot offer flexible packaging when the external identifier for a bag is used as identifier for a file, since this means that a bag can only include one file.

**Table 1. Package formats fulfilment of requirements**

| Requirements \ Formats | AFF | ARC | BagIt | METS | RAR | TAR | WARC | ZIP |
|---|---|---|---|---|---|---|---|---|
| **1. Platform independent** | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **2. Flexible packaging** | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| **3. Supports update packages** | No | No | No | Almost | No | No | Yes | No |
| **4. Standardised** | Little | No | So-so | Yes | No | Yes | Yes | Little |
| **5. Open** | Yes | Yes | Yes | Yes | No | Yes | Yes | Almost |
| **6. Easily understandable** | So-so | So-so | So-so | Almost | No | Little | Yes | Little |
| **7. Widely used in bit repositories** | No | So-so | Almost | So-so | Little | Yes | Almost | So-so |
| **8. Tools available** | So-so | Yes | Yes | So-so | Yes | Yes | So-so | Yes |
| **9. Include files unchanged** | Yes | Yes | Yes | No | No | Yes | Yes | Yes |
| **10. Identifiers for files** | Yes | So-so | So-so | Yes | No | No | Yes | No |

## 3.3 Suggested choice of WARC

The requirements ranked in table 1 should not be equally weighted. First of all the importance of long term preservation is regarded as highest. Secondly, there are requirements that become less important, if other requirements are given high score. For instance, it may not be important that a format is *Standardised*, in case the format has high scores on *Easily understandable*, *Open* and *Widely used*. Such a format may have a higher chance of surviving as a de facto standard, than another standardised format which is neither *Easily understandable* nor *Widely used*. Similar for tooling, a format that is *Open* and *Widely used* is quite likely to get *Tools available* in a relatively short time.

The final suggestion of WARC is therefore based on analysis that takes such considerations into account, and using exclusion of formats by comparison between the formats.

**ARC** can be ruled out, since it is a much more primitive and immature package format than WARC, thus arguments for choosing ARC will also be arguments for choosing WARC, but WARC has more benefits than ARC.

**METS** and **XFDU** can be ruled out, since they are XML based which cannot support proper inclusion of files, which is crucial and thus a mandatory requirement for the long term preservation.

**RAR** is ruled out since it can only offer compressed packaging which cannot be accepted for *all* long term preservation.

If the requirement to assign identifiers for files is considered crucial, then the **TAR** and **ZIP** formats are best considered in connection with the BagIt format. From table 1 it is evident that the TAR format better fulfils other requirements, since it has the same score or better score than the ZIP format for the same requirements.

The only real problem with **BagIt** is that it only can have one external identifier assigned to a package, which is probably due to the fact that it is designed as an exchange format. This fact means that settling for BagIt would limit the possibilities of how to make packages, since use of external identifiers for identifiers means that a bag can only have one file. However, it only has low ranking of requirements that are considered less important for long term preservation, and it is therefore worthwhile to consider this format. However, besides BagIt, there will have to be a decision on whether it should build on TAR or ZIP.

The **WARC** format is a candidate since it can support all requirements, although it is not widely used yet (at least as package format for all types of digital material), and there is no stable tool package to support it. However, there are a lot of indications that this will change to the better, since web archives will start to use WARC instead of ARC. Furthermore, using WARC for other than web material is not entirely new. For instance the German Kopal project is today working towards packaging all types of materials in WARC when sent to bit preservation [21] (using Private LOCKSS Networks [19]).

Finally the AFF format could be a candidate, but compared to BagIt and WARC, it loses on the fact that there is limited experience in use as a general package format, and is not widely used. As presented in the [10] WARC only lacks the ability to represent file system structure or the file system characteristics in order to meet requirements for forensic data. However, in the preservation perspective taken in this paper, this is not crucial,

since such metadata can just as well be part of the packed metadata.

The two most relevant alternatives found in the analysis are therefore WARC and BagIt based on TAR.

The only requirement where WARC scores lower than BagIt is the same requirement as the lowest score for WARC, namely: *Tools available*. This means that there may be a risk that local investments must be made for tools using WARC. However, the interest in using WARC for web archiving indicates that a community for tool development exists and tools probably will emerge soon.

The two formats have the same score *Widely used*, but for different reasons. Although BagIt is designed as an exchange format, it is also used for repository material. WARC on the other hand is mostly used for web archive material, or is most likely to be used in most future web archives. The risk that they may not go for the WARC format after all is quite slim, since WARC is now both the only formally standardised format for web archiving, but also the best alternative, since it is developed based on previous experiences with web archiving formats like ARC.

Great advantages with WARC compared to BagIt are that it can represent *Identifiers for files* easily, and a WARC package is in easily understandable text form. On the other hand BagIt can only represent one external identifier per bag and interpretation relies on knowledge of both BagIt and TAR.

The restraints on how to use external identifiers in BagIt also mean that the WARC format is best with regard to flexible packaging. This enables the possibility of choosing to put metadata for files in the same package as the file, or even more objects in the same package. As the size of packages can have impact on different resource issues the flexibility in settling for policies in using WARC can affect optimization resource use.

Finally the WARC format is the only format of the mentioned ones[23], where it is possible to define update records directly. This is not the most crucial requirement, but it can help to optimise preservation costs, if the risk analysis from bit preservation can allow preservation of updates as an alternative way of preserving a representation.

Besides the advantages that WARC have considering the requirements, WARC also has an extra advantage for institutions with web archives using WARC: The institution will only need skills concerning WARC as package format for all preserved data. This is for instance the case for The Royal Library of Denmark. It should however be noted that the way WARC is used for web archives may be more advanced than the way WARC is used for other materials. Still it is a great advantage not to need skills for more package formats.

A discussable advantage of WARC is that it does not rely on assumptions of having folder and file structures. As expressed in "Cedars Guide to: Digital Preservation Strategies"[24].

"The UNIX format known as tar (originally standing for tape archive) is used by Cedars as the preservation byte-stream for such cases, because it is publicly documented, and there exists public domain software for writing and reading data in such a format. Another institution may choose to use a different format for mapping the original file tree into a byte stream. Whatever format is chosen, it must enable a subsequent recreation of a file system that operates in the same way as the original. Thus the files system should be converted to a byte-stream for preservation by use of tar or other suitable program."

In other words TAR does have assumption of file and folder structures as the basis for unpacking the TAR file. Whether this will exist in 100 years can only be a guess, thus there will be different opinions on whether risk of losing the basis for unpacking TAR files should be included in a risk analysis as basis for choosing a package format.

## 4. DISCUSSION

It could be argued that this paper should have included a more complete list of formats that can be used for packaging data that are to be bit preserved. However, most other alternatives are less known formats, commercial formats or formats designed for a specific purpose. Thus such formats would most likely be eliminated on requirements of being open, standardised and widely used.

This paper has only included the most relevant requirements for preservation of general digital materials. There can be supplementary requirements for e.g. how the format supports availability of data. Such requirements are described in the literature consisting of guidelines, reports and papers [2,10,12,13,14,18,22].

The requirement of expressing *Identifiers for files* is crucial for the choice of WARC in the present presentation. Therefore there may be cases where such an analysis will not lead to the same result. This would for instance be the case where this requirement is seen as less important, due to e.g. relying on a bit repository to keep track of the identifier, having few formats where risk of losing embedded identifiers is seen as unimportant, or risk related to having identifiers as part of the file name is considered minor.

Another example, where analysis of choice for package format is different, is the package format for forensic digital material as given in [10]. This is due to the fact that the requirements and focus are different. It may be that the choice of package format will be different for different types of digital material, e.g. forensic and other digital material. However, it should be noted that there are no limitations in WARC to include AFF packages. This could be desirable in the case of the benefits of a general package format in a bit repository, e.g. in order to have similar access to all packages. However it can also be considered more beneficial to have several package formats, since overhead in unpacking, and possibly impact of access time of the data can be avoided. Likewise, there could be other specific digital material that needed specific considerations, e.g. specific scientific data.

The packaging for bit preservation may not be optimal for the way digital material is e.g. disseminated. The focus is on preservation. Thus the focus regarding availability is that it will be possible to reproduce digital material and identifiers, solely based on the preserved packages. This means that additional analysis will be required for cases where there are specific

[23] Other formats supports update specification, e.g. VCDIFF (http://tools.ietf.org/html/rfc3284), but these are typically not suited as general package formats

[24] See http://www.imaginar.org/dppd/DPPD/146%20pp%20Digital%20Preservation%20Strategies.pdf

requirements to access time that are more important than preservation requirements.

## 5. CONCLUSION

This paper found the best suited format for long term preservation of varied digital materials is WARC. However, the value of the analysis depends on whether the presented requirements are seen as the most important requirements for the digital preservation of the material, and whether there are other requirements to be included.

Compared to most other formats, the WARC format is strong as a preservation packaging format in general, especially regarding issues of: applying identifiers to bit-sequences/files, being easily understandable and being one of the few formally standardised formats. Furthermore the WARC format is the only format among the listed formats that is extendible with record definition for update records, which can give economical benefits for preserving changing materials.

The only point where the WARC format does not have the top score is how widely used the format is, and how well it is supported by tools. However, the lower score concerning 'widely used' is based on the fact that it is mostly used within web archiving, although there are no restrictions or overhead in using the WARC format for other types of digital archiving. Regarding tool support, the increasing use of the WARC format gives reasons to believe that this will change to the better.

## 6. REFERENCES


[1] CCSDS (Consultative Committee for Space Data Systems): *XML Formatted data Unit (XFDU) structure and Construction Rules*, CCSDS 661.0-B-1, Blue book, http://public.ccsds.org/publications/archive/661x0b1.pdf, 2008.

[2] Christensen, S. S. *Archival Data Format Requirements*, The Royal Library, Copenhagen, Denmark, The State and University Library, Aarhus, Denmark. http://netarchive.dk/publikationer/Archival_format_requirements-2004.pdf, 2004.

[3] Cohen, M., Garfinkel, S., Schatz, B. Extending the advanced forensic format to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflow, In: Digital Investigation - The International Journal of Digital Forensics & Incident Response, no. 6, pp. 57-68, 2009.

[4] Cramer, T., Kott, K. Designing and Implementing Second Generation Digital Preservation Services: A Scalable Model for the Stanford Digital Repository, In: *D-Lib Magazine, vol. 16, no. 9/10*. 2010.

[5] Enders, M.: A METS Based Information Package for Long Term Accessibility of Web Archives, In: *Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria,* 2010.

[6] Gillesse, R., Rog, J., Verheusen, A. Life Beyond Uncompressed TIFF: Alternative File Formats for Storage of Master Image File.' In: *Proceedings of the IS&T Archiving Conference*, Bern, Switzerland, 2008.

[7] ISO 28500:2009, *Information and documentation -- WARC file format*, retrievable via http://www.iso.org/iso/iso_catalogue.htm, 2009.

[8] ISO 9660:1988, ECMA-119, *Optical media disk imaging format*,, retrievable via http://www.iso.org/iso/iso_catalogue.htm, 2010.

[9] Kejser, U.B.: Preservation copying of endangered historic negative collections, In: Proceedings of the IS&T Archiving Conference, Bern, Switzerland, pp. 177-182, 2008.

[10] Kim, Y. Digital forensics formats: seeking a digital preservation storage format for web archiving. In *Proceedings of 7th International Digital Curation Conference*, Bristol, United Kingdom, 2011.

[11] Lavoie, B., Dempsey, L.: Thirteen Ways of Looking at ... Digital Preservation, In: *D-Lib Magazine vol. 10 no. 7/8*, 2004.

[12] *Local Digital Format Registry (LDFR), File Format Guidelines for Preservation and Long-term Access,* Version 1.0, Library and Archives Canada, http://www.collectionscanada.gc.ca/digital-initiatives/012018-2210-e.html, 2010.

[13] Library of Congress. Sustainability of Digital Formats: Planning for Library of Congress Collections, http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

[14] National Archives (UK). *Selecting File Formats for Long-Term Preservation,*http://www.nationalarchives.gov.uk/documents/selecting_file_formats.rtf, 2003.

[15] Oury, C., Peyrard, S.: From the World Wide Web to digital library stacks: preserving the French web archives In: *Proceedings iPRES 2011, 8th International Conference on Preservation of Digital Objects, Singapore, Singapore, 2011*

[16] Phillip Katz, *Computer Software Pioneer, 37*, In: The New York Times, May 1st 2000, http://www.nytimes.com/2000/05/01/us/phillip-katz-computer-software-pioneer-37.html, 2000.

[17] PREMIS Editorial Committee, *PREMIS Data Dictionary for Preservation Metadata*, version 2.1, http://www.loc.gov/standards/premis/v2/premis-2-1.pdf, 2011.

[18] Rauch, C., Krottmaier, H. Tochtermann, K.: File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats, In: *Proceedings ELPUB2007 Conference on Electronic Publishing, Vienna, Austria, 2007*.

[19] Reich, V., Rosenthal, D. S. H.: Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks, In: *Library Trends, vol. 57, no. 3, pp. 461-475*, 2009

[20] Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V., Morabito, S.: Requirements for Digital Preservation Systems, A Bottom-Up Approach, In: *D-Lib Magazine, vol. 11, no. 11,* 2005.

[21] Seadle, M.: Archiving in the networked world: LOCKSS and national hosting, In: Library Hi Tech, vol. 28, Issue 4, pp. 710-717 (2010)

[22] The InterPARES 2. *Project. General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation*, http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf, 2006.

[23] WARC, Web ARChive file format, in: Sustainability of Digital Formats Planning for Library of Congress Collections, http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

[24] Wright, R., Miller, A., Addis, M.: The Significance of Storage in the "Cost of Risk" of Digital Preservation, In: *The International Journal of Digital Curation, vol. 4, issue 3*, pp. 105-122, 2009.

[25] Zierau, E. *A Holistic Approach to Bit Preservation*, Doctoral Thesis, University of Copenhagen, http://www.diku.dk/research/phd-studiet/phd/thesis_20111215.pdf/, 2011.