Spezifikation BagIt DLA Netzliteratur

Das vorliegende Papier beschreibt Bags. Bags stellen die Archivalien dar, die das Deutsche Literaturarchiv Marbach (nachfolgend DLA) im Bereich Netzliteratur an das Bibliotheksservice-Zentrum Baden-Württemberg (nachfolgend BSZ) übertragen wird.

Diese Spezifikation folgt dabei dem Draft *The BagIt File Packaging Format* der Internet Engineering Task Force (nachfolgend IETF) in Version 0.97[1].

Struktur einer Bag

Bags fassen unterschiedliche digitale Objekte in einer Hierarchie zusammen.
Eine Bag ist ein Ordner, der eindeutig definierte Elemente enthalten muss sowie weitere arbiträre Daten enthalten kann.

Wurzel

Die oberste Ebene einer Bag ist ein Verzeichnis. Die Benennung dieses Verzeichnisses erfolgt einheitlich nach dem folgenden Muster:

MAB026_[UUID]_JJJJMMDD

MAB026 bezeichnet ein Feld im Verbundkatalog und garantiert eine eindeutige Zuordnung zu einer Titelaufnahme. Nach dieser Bezeichnung folgt ein Unterstrich. Diesem kann eine UUID folgen. Den letzten Teil des Namens bildet das Datum der Erstellung der Bag im Basisformat nach ISO 8601:2004.

Beispiele:

1. bsz396664105_5c6b3e91-471f-4504-9d57-b9d088093b77_20140319/
2. bsz396664105_20140319/

Obligatorische Elemente

In einer Bag müssen die folgenden Dateien enthalten sein. Dies stellt eine Erweiterung, bzw. Konkretisierung der Anforderungen nach dem zu Grunde liegenden Entwurf der IETF dar. Unter der Benennung der Elemente findet sich jeweils eine Erläuterung sowie ein Beispiel.

1. bag-info.txt

   Die Datei bag-info.txt beinhaltet den **Namen des Programms**, mit dem die Bag erstellt wurde sowie das **Erstellungsdatum**. Außerdem findet sich in ihr die **Payload-Oxum**. Die Payload-Oxum gibt Auskunft darüber, wie viele Bytes insgesamt auf wie viele Dateien insgesamt im Verzeichnis data/ verteilt sind. Das DLA ergänzt diese obligatorischen Informationen um den Namen der **sendenden Organisation** (SOURCE_ORGANIZATION) sowie den Namen eines/einer **Verantwortlichen** (Contact-Name).

   Beispiel:

   Bag-Software-Agent: bagit.py <http://github.com/edsu/bagit>
   Bagit-Date: 2014-02-10
   Contact-Name: Steffen Fritz
   Payload-Oxum: 268597:12
   SOURCE_ORGANIZATION: Deutsches Literaturarchiv Marbach

[1] vgl. http://tools.ietf.org/html/draft-kunze-bagit, zugegriffen am 17.02.2014

2. bagit.txt

   Die Datei bagit.txt beinhaltet die **Version der** zu Grunde liegenden **BagIt-Spezifikation** sowie das **Character-Encoding** der Tag-Files, das immer „UTF-8“ sein muss.

   Beispiel:

```
BagIt-Version: 0.97
Tag-File-Character-Encoding: UTF-8
```

3. manifest-HASHFUNKTION.txt

   Die Datei manifest-HASHFUNKTION.txt **listet alle Dateien sowie deren Checksummen** auf, die sich im Verzeichnis data/ befinden. Der Name der verwendeten Hashfunktion muss im Dateinamen genannt sein. Das DLA wird die Hashfunktion Secure Hash Algorithm mit einer Länge von 512 Bit verwenden, nachfolgend sha512 genannt. Daher befindet sich in jeder Bag eine Datei mit der Bezeichnung **manifest-sha512.txt**.

   Beispiel:

```
63984ff676545hhafe198(...)4637b567c21 data/metadata.xml
f83f16763457eeafe49851(...)627e5eefc94 data/screenshot_00.jpg
ee257633371aaf34785af (...)1e151337c1 data/screenshot_00.tiff
fa972fe6e66974aafe1934(...)433e567d37  data/ampoffcom_20140101.warc.gz
```

4. data/

   Das Verzeichnis data/ bildet die zweite Ebene einer Bag. Es ist das Payload-Verzeichnis. In ihm befinden sich die archivierten Daten.

   Beispiel:

```
metadata.xml
screenshot_00.jpg
screenshot_00.tif
ampoffcom_20140101.war.gz
```

5. metadata.xml

   Die Datei metadata.xml beinhaltet alle Metadaten, die hinsichtlich des Werks erfasst wurden.

6. screenshot_uv.jpg

   Jedem Archiv wird mindestens ein Bildschirmfoto des Werks im JPG-Format beigelegt. Sollte dies nicht möglich sein, wird ein Platzhalterfoto beigelegt. Der Bezeichnung *screenshot* folgt, getrennt durch einen Unterstrich, eine zweistellige laufende Nummer, die es ermöglicht, mehrere Screenshots beizulegen.

7. screenshot_uv.tif

   Jedem Archiv wird mindestens ein Bildschirmfoto des Werks im TIF-Format beigelegt. Sollte dies nicht möglich sein, wird ein Platzhalterfoto beigelegt. Der Bezeichnung *screenshot* folgt, getrennt durch einen Unterstrich, eine zweistellige laufende Nummer, die es ermöglicht, mehrere Screenshots beizulegen.

Sequentielle Speicherung

Die Bag wird mittels **tar** zu einer Datei zusammengefasst. Daraus ergibt sich ein Archivname der folgenden Form:

MAB026_[UUID]_JJJJMMDD.tar

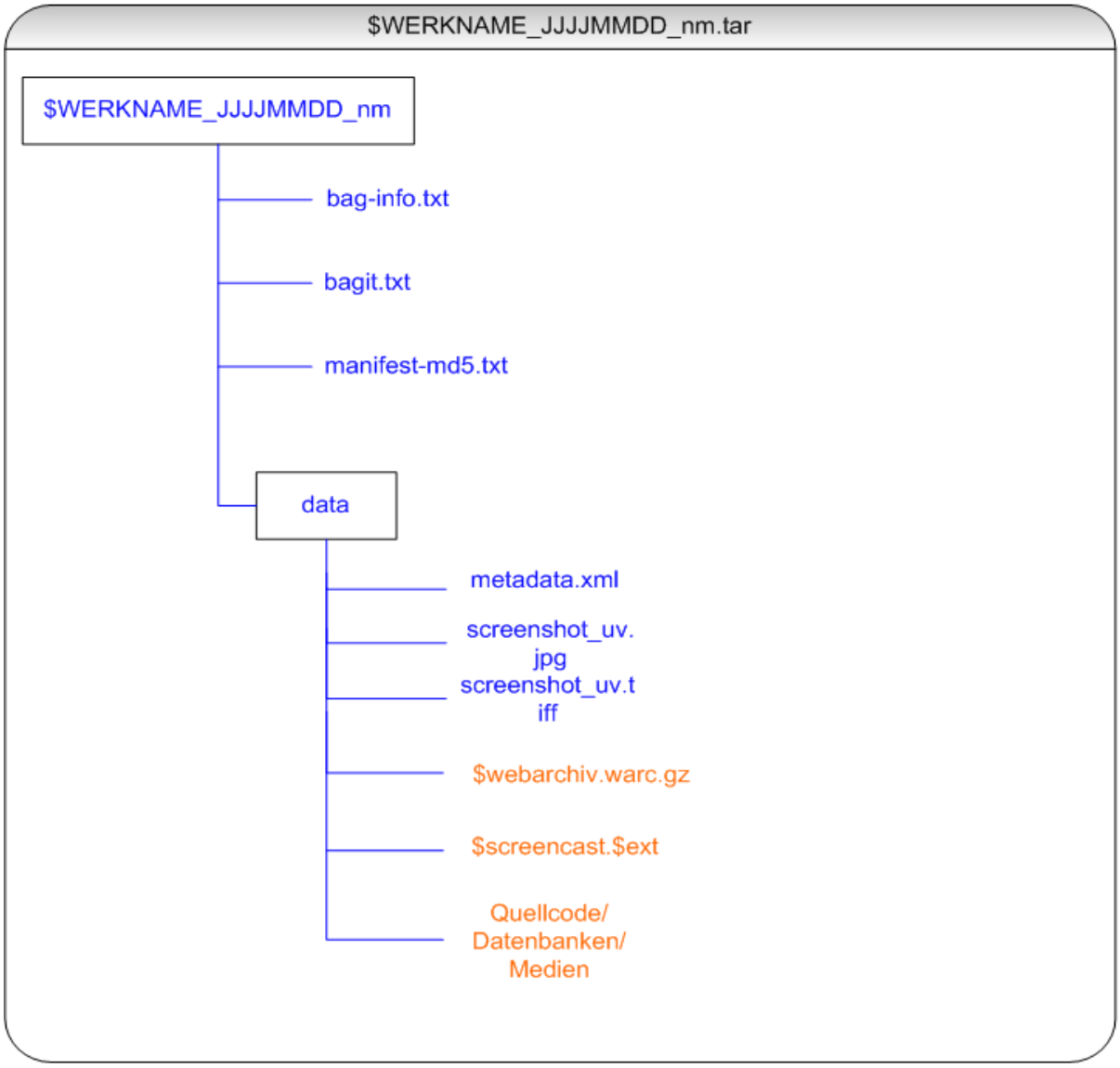
Diese Datei wird anschließend mittels gzip komprimiert, ein entsprechendes Suffix wird angehängt. Daraus ergibt sich der Archivname

MAB026_[UUID]_JJJJMMDD.tar.gz

Schematische Darstellung einer Bag

Titel: Aufbau BagIt DLA Netzliteratur
Autor: Steffen Fritz

Datum: 31.03.2014
Dateiname: BagIt_Uebersicht_20140331.03a.vsd



bag-info.txt	obligatorisch	Informationen zu BagIt-Software-Agent, Datum und Oxum (Byte.Files)
bagit.txt	obligatorisch	BagIt-Version und Tag-File-Encoding
manifest-md5.txt	obligatorisch	Dateiliste mit Hashes. Aufbau: HASH PFAD/DATEINAME
data/	obligatorisch	Payload-Ordner
metadata.xml	obligatorisch	Metadaten
screenshot_uv.jpg	obligatorisch	Screenshot des Werkes. Falls nicht möglich, wird ein Platzhalter beigelegt
screenshot_uv.tiff	obligatorisch	Screenshot des Werkes. Falls nicht möglich, wird ein Platzhalter beigelegt
$webarchiv.warc.gz		
$screencast.$ext		
Quellcode etc.		

The BagIt File Packaging Format v0.97, Stand: 28. Januar 2014,
vgl. http://tools.ietf.org/html/draft-kunze-bagit-10