

Preservation Planning, Preservation Actions

1. Definition der Archivformate

Die Definition der Archivformate geschieht in der Konfigurationsdatei „migration-config.xml“ von docuteam feeder. In dieser umfangreichen Datei wird für jedes Ausgangsformat ein Archivformat hinterlegt. Diese Konfiguration kommt zur Anwendung unabhängig davon, ob es sich um einen initialen Ingest oder um die Migration von Dateien handelt, die bereits im Archiv liegen. Die folgende Abbildung zeigt einen Ausschnitt aus dieser Konfigurationsdatei. Für jedes Quellformat wird sowohl ein Zielformat definiert, wie auch die für die Migration verantwortliche Applikation hinterlegt.

```
c:\> MD5 PROCKING --S>

c!-- Word for Windows -->
cpuid name=fat/37' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/38' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/39' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/40' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/41' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/503' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/597' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/598' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />

c!-- Word for H3-200 -->
cpuid name=a-fat/373' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/374' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/375' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/376' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />

c!-- Word for Mac -->
cpuid name=a-fat/2' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/139' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />

c!-- Wordperfect -->
cpuid name=a-fat/42' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/43' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/44' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/203' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/303' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/394' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=a-fat/403' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />

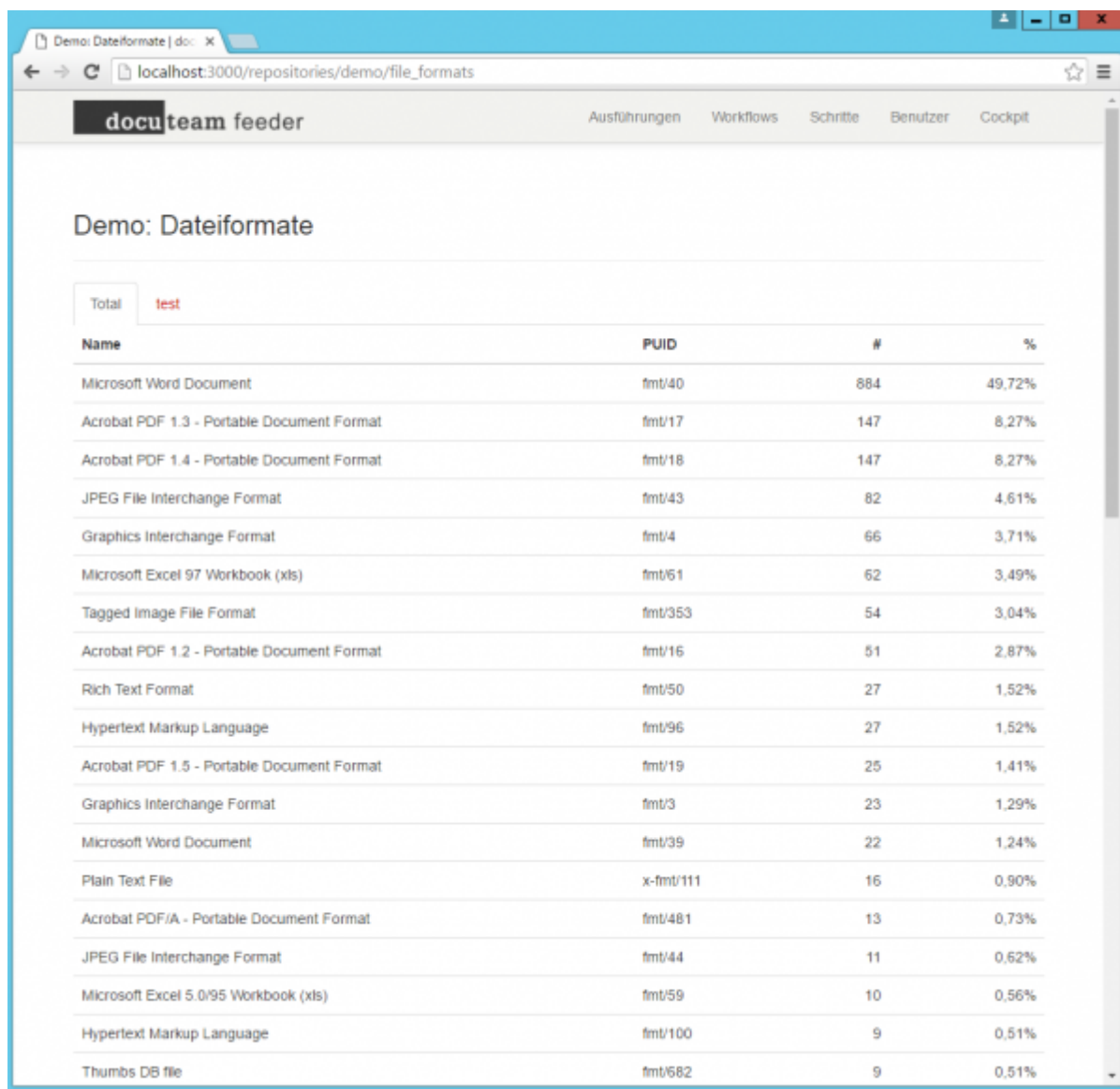
mimeType name=application/vnd.openxmlformats-officedocument.wordprocessingml.template' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
mimeType name=application/vnd.ms-word.template.maindocbled.12' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
mimeType name=application/vnd.ms-word.document.macrohtml.12' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
mimeType name=application/vnd.wordperftot' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
mimeType name=application/msword' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
extension name=.doc' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
extension name=.docm' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
extension name=.docx' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />

c!-- Rich Text -->
cpuid name=fat/45' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/46' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/49' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/40' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/48' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/50' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/51' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/52' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
cpuid name=fat/53' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />

mimeType name=application/rtf' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
mimeType name=text/rtf' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
extension name=.rtf' applicationID=6' targetExtension=.pdf' targetFromme=fat/479,fat/480,fat/481' />
```

2. Grundlagen für das Preservation Planning ermitteln

Die Grundlagen für das Preservation Planning werden im Cockpit von docuteam feeder über den Bericht „Dateiformate“ ermittelt. Der Bericht listet – geordnet nach Häufigkeit – alle im Archiv vorhandenen Dateiformate auf und zeigt auch ihren Pronom-Identifizier (PUID) an. [Pronom](#) ist die Fileregistry des britischen Nationalarchivs und erlaubt es, Dateiformate und Unterversionen davon eindeutig zu identifizieren und zu referenzieren. Aufgrund dieses Berichts wird der Entscheid gefällt, ob und welche Dateiformate in neue Zielformate migriert werden sollen.



The screenshot shows the 'Demo: Dateiformate' page in the Docuteam Feeder application. The browser address bar shows 'localhost:3000/repositories/demo/file_formats'. The page has a navigation bar with 'Ausführungen', 'Workflows', 'Schritte', 'Benutzer', and 'Cockpit'. Below the title, there are tabs for 'Total' and 'test'. The main content is a table with four columns: 'Name', 'PUID', '#', and '%'. The table lists various file formats and their corresponding statistics.

Name	PUID	#	%
Microsoft Word Document	fmt/40	884	49,72%
Acrobat PDF 1.3 - Portable Document Format	fmt/17	147	8,27%
Acrobat PDF 1.4 - Portable Document Format	fmt/18	147	8,27%
JPEG File Interchange Format	fmt/43	82	4,61%
Graphics Interchange Format	fmt/4	66	3,71%
Microsoft Excel 97 Workbook (xls)	fmt/51	62	3,49%
Tagged Image File Format	fmt/353	54	3,04%
Acrobat PDF 1.2 - Portable Document Format	fmt/16	51	2,87%
Rich Text Format	fmt/50	27	1,52%
Hypertext Markup Language	fmt/96	27	1,52%
Acrobat PDF 1.5 - Portable Document Format	fmt/19	25	1,41%
Graphics Interchange Format	fmt/3	23	1,29%
Microsoft Word Document	fmt/39	22	1,24%
Plain Text File	x-fmt/111	16	0,90%
Acrobat PDF/A - Portable Document Format	fmt/481	13	0,73%
JPEG File Interchange Format	fmt/44	11	0,62%
Microsoft Excel 5.0/95 Workbook (xls)	fmt/59	10	0,56%
Hypertext Markup Language	fmt/100	9	0,51%
Thumbs DB file	fmt/682	9	0,51%

3. Extraktion der zu migrierenden Dateien

Preservation Actions werden immer pro Dateiformat durchgeführt. Beispiel: Im Archiv liegt verschiedene JPEG-Dateien, die im Rahmen einer Preservation Action nach Tiff migriert werden sollen. In migration.config.xml ist Tiff als Archivformat von JPEG hinterlegt.

JPEG (Pronom fmt/43) -> Tiff (Pronom fmt/353)

In docuteam feeder werden mit Hilfe des Workflows „Preservation: Deliver DIP by PUID“ alle JPEG-Dateien extrahiert. Der Workflow wird gestartet und als Parameter wird im Feld „Manuelle Eingabe“ der PUID des zu migrierenden Dateiformats angegeben. **Wichtig:** der PUID muss in Anführungszeichen gesetzt werden, also **„fmt/43“** und nicht nur **fmt/43**.

The screenshot shows the 'docuteam feeder' web interface. At the top, there are navigation tabs: 'Ausführungen', 'Workflows', 'Schritte', 'Benutzer', and 'Cockpit'. The main heading is 'Neue Ausführung für Preservation: Deliver DIP by PUID'. Below this, there is a form with the following elements:

- A label 'Objekt-ID' followed by a dropdown menu labeled 'SIP'.
- A label 'oder'.
- A label 'Manuelle Eingabe' followed by a text input field containing the value 'fnt/43'.
- A button labeled 'Ausführung erstellen'.

Die JPEG-Dateien werden je als eigene Informationspakete im Ordner 1_hotfolder_preservation auf der Workbench gespeichert:

The screenshot shows a Windows File Explorer window titled '1_hotfolder_preservation'. The address bar indicates the path: 'This PC > Local Disk (C:) > workbench > 1_hotfolder_preservation'. The left sidebar shows the 'Local Disk (C:)' selected. The main pane displays a list of files with the following columns: Name, Date modified, Type, and Size.

Name	Date modified	Type	Size
test_133.zip	13.07.2016 10:52	Compressed (zipp...	37 KB
test_134.zip	13.07.2016 10:53	Compressed (zipp...	33 KB
test_136.zip	13.07.2016 10:52	Compressed (zipp...	26 KB
test_152.zip	13.07.2016 10:52	Compressed (zipp...	37 KB
test_153.zip	13.07.2016 10:52	Compressed (zipp...	33 KB
test_209.zip	13.07.2016 10:52	Compressed (zipp...	22 KB
test_210.zip	13.07.2016 10:52	Compressed (zipp...	30 KB
test_211.zip	13.07.2016 10:52	Compressed (zipp...	94 KB
test_241.zip	13.07.2016 10:52	Compressed (zipp...	10 KB
test_247.zip	13.07.2016 10:52	Compressed (zipp...	9 KB
test_248.zip	13.07.2016 10:52	Compressed (zipp...	9 KB
test_250.zip	13.07.2016 10:52	Compressed (zipp...	10 KB
test_251.zip	13.07.2016 10:52	Compressed (zipp...	10 KB
test_252.zip	13.07.2016 10:52	Compressed (zipp...	11 KB
test_254.zip	13.07.2016 10:53	Compressed (zipp...	8 KB
test_298.zip	13.07.2016 10:52	Compressed (zipp...	20 KB

3. Durchführung der Preservation Action

Die Preservation Action kann für Einzeldateien manuell über den Workflow „Update Fedora Objects“ durchgeführt werden. In der manuellen Eingabe wird dem Workflow der Name des zu migrierenden Pakets angegeben. Die manuell ausgeführte Preservation Action dient primär dem Testbetrieb, bei einer größeren Menge von zu migrierenden Dateien wird der Workspace-Ordner „1_hotfolder_preservation“ durch einen observer überwacht und sobald dort neue Pakete eintreffen werden sie automatisiert in das neue Zielformat migriert.

docuteam feeder Ausführungen Workflows Schritte Benutzer Cockpit

Neue Ausführung für Preservation: Update Fedora Objects

Objekt-ID

SIP

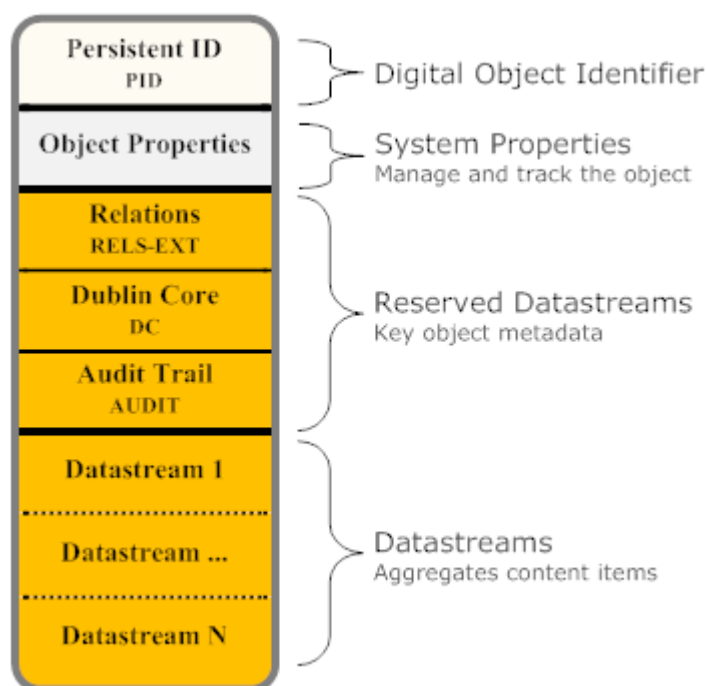
oder

Manuelle Eingabe

Ausführung erstellen

Angemeldet als admin | [Logout](#) | [Dokumentation](#) | [Hilfe](#)

Das [Objektmodell](#) von Fedora Commons sieht vor, dass Dateien in sogenannten „Datastreams“ gespeichert werden und dass diese Datastreams versioniert werden können. Docuteam legt für jede Datei ein eigenes Fedora-Objekt an und verknüpft die Objekte über Relationen (beschrieben im RELS-EXT Datastream). Wenn nun eine Datei in ein neues Format migriert wird, dann wird der entsprechende Datastream nicht einfach überschrieben, sondern es wird eine neue Version davon angelegt. Dadurch bleibt der PID der Datei weiterhin gültig. Die Ausgangsdatei wird bei einer Preservation Action nicht gelöscht. Falls also im Migrationsprozess ein Fehler auftritt, kann auch zukünftig jederzeit auf die Ursprungsdatei zurückgegriffen werden. Während der Preservation Action werden zur Vorgangsdokumentation auch neue PREMIS-Metadaten geschrieben. Diese liegen in Fedora ebenfalls in einem eigenen Datastream und auch dieser wird versioniert.



From:
<https://wiki.docuteam.ch/> - **docuteam wiki**

Permanent link:
https://wiki.docuteam.ch/doku.php?id=docuteam:preservation_330&rev=1499774206

Last update: **2019/01/07 11:16**

