



Newspaper Digitisation

Policy driven validation of JPEG 2000 files based on Jpylyzer

Rune Bruun Ferneke-Nielsen

State and University Library, Denmark

SCAPE Information Day

State and University Library, Denmark, June 25th 2014

Agenda

- Newspaper Digitisation Project
- User Story & Experiment
- Results

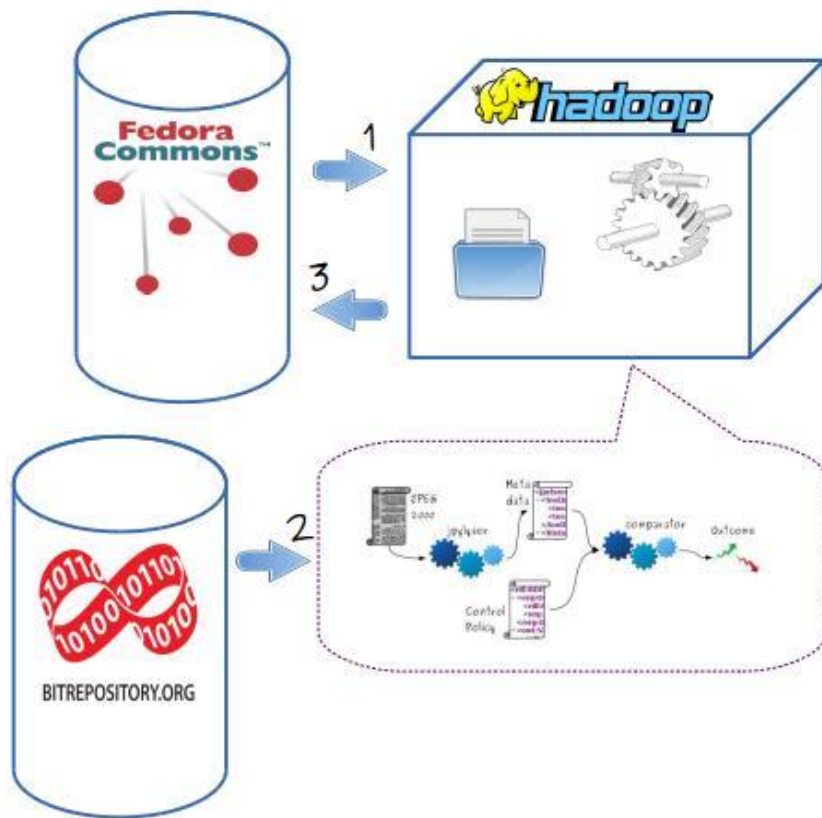
- Preservation of Danish cultural heritage
 - 32 million pages scanned from microfilm
 - Quality assurance of digitised pages
 - Online access through Mediestream
-
- Project Period: 2013 - 2016
 - State and University Library, Denmark
 - Ninestars Information Technologies Ltd, India

Validation of Archival Content Against an Institutional Policy

As a memory institution, we want

- content in our repositories to conform to the corresponding file format specification
- the file format profile to conform to our institutional policies

So that our content - existing as well as future - always has the appropriate quality as specified by the file format specification and our institutional policies.



1. Extracting metadata from Fedora-based repository
2. Performing quality assurance on Hadoop platform
3. Storing metadata into Fedora-based repository

- Stager component input
- Using Stager component
 - Reading DOMS objects
- Using sequence file
 - Sequence files are flat files consisting of key/value pairs

```
uuid:723110d9-c76d-4e1b-870d-4968c3ebdf51  
uuid:1c0194a3-c5af-4b40-b140-5ac64cfa43af  
uuid:ef88a6b3-2ce3-4dd5-b0c7-8bebe1656aa5  
uuid:d7183549-f108-40a3-81a8-704114667174  
uuid:548a7a25-82d6-47b5-9e14-eae3adabd423
```

key	value
uuid:723110d9-c76d-4e1b-870d-4968c3ebdf51	<metadata>
uuid:1c0194a3-c5af-4b40-b140-5ac64cfa43af	<metadata>
uuid:ef88a6b3-2ce3-4dd5-b0c7-8bebe1656aa5	<metadata>
uuid:d7183549-f108-40a3-81a8-704114667174	<metadata>
uuid:548a7a25-82d6-47b5-9e14-eae3adabd423	<metadata>

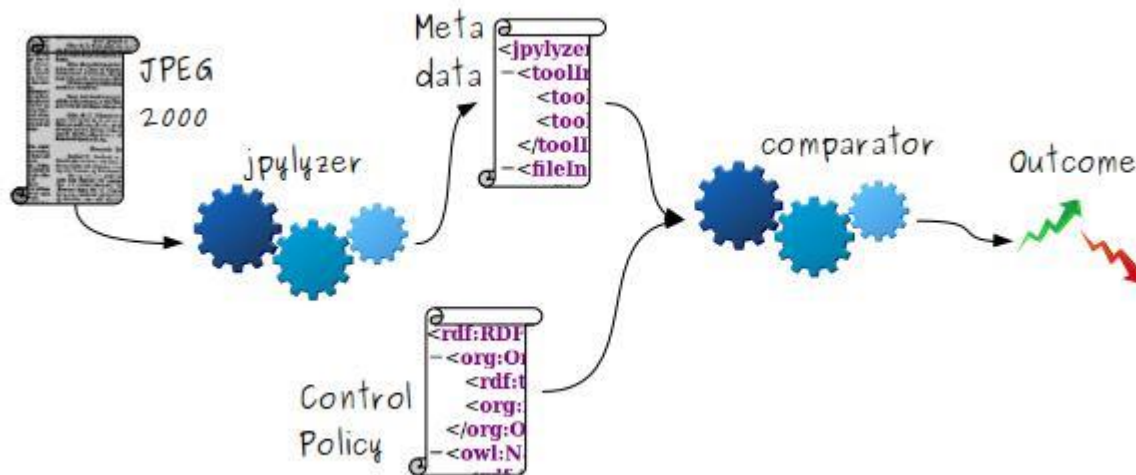
METS Document from DOMS

```

1 <?xml version='1.0' encoding='UTF-8'?>
2 <mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:xlink="http://www.w3.org/1999/xlink"
3   xmlns:scape="http://scape-project.eu/model" xmlns:dc="http://purl.org/dc/elements/1.1/"
4   xmlns:premis="info:lc/xmlns/premis-v2" xmlns:textmd="info:lc/xmlns/textmd-v3"
5   xmlns:fits="http://hul.harvard.edu/ois/xml/ns/fits/fits_output"
6   xmlns:ns9="http://www.loc.gov/mix/v20" xmlns:gsb="http://books.google.com/gbs"
7   xmlns:vmd="http://www.loc.gov/videoMD/" xmlns:ns12="http://www.loc.gov/audioMD/"
8   xmlns:marc="http://www.loc.gov/MARC21/slim" ID="uuid:ba6865e5-7682-4cc0-8e80-0f1ccac3c027"
9   OBJID="ba6865e5-7682-4cc0-8e80-0f1ccac3c027" PROFILE="scape">
10   <mets:metsHdr RECORDSTATUS="NEW" />
11   <mets:dmdSec ID="DMD-3f4117c1-dedd-4598-9195-f736b78705eb">
12     <mets:mdWrap MDTYPE="Other">
13       <mets:xmlData />
14     </mets:mdWrap>
15   </mets:dmdSec>
16   <mets:dmdSec ID="DMD-0273b15f-a3f5-4484-b773-357551800ffa">
17     <mets:mdWrap MDTYPE="OTHER">
18       <mets:xmlData>
19         <scape:versionMD version-number="1" />
20       </mets:xmlData>
21     </mets:mdWrap>
22   </mets:dmdSec>
23   <mets:amdSec>
24     <mets:techMD
25       ID="TMD-ba6865e5-7682-4cc0-8e80-0f1ccac3c028-SCAPE_REPRESENTATION_TECHNICAL">
26       <mets:mdWrap MDTYPE="OTHER">
27         <mets:xmlData />
28       </mets:mdWrap>
29     </mets:techMD>
30     <mets:techMD
31       ID="TMD-ba6865e5-7682-4cc0-8e80-0f1ccac3c029-SCAPE_FILE_TECHNICAL">
32       <mets:mdWrap MDTYPE="OTHER">
33         <mets:xmlData />
34       </mets:mdWrap>
35     </mets:techMD>
36   </mets:amdSec>
37   <mets:fileSec>
38     <mets:fileGrp>
39       <mets:file ID="ba6865e5-7682-4cc0-8e80-0f1ccac3c029">
40         SEQ="0"
41         ADMID="TMD-ba6865e5-7682-4cc0-8e80-0f1ccac3c029-SCAPE_FILE_TECHNICAL"
42         MIMETYPE="image/jp2">
43         <mets:FLocat
44           xlink:href="src/test/resources/sample/adresseavisen1759-1795-06-13-01-0006.jp2"
45           xlink:title="adresseavisen1759-1795-06-13-01-0006.jp2" LOCTYPE="URL" />
46         </mets:file>
47       </mets:fileGrp>
48     </mets:fileSec>
49     <mets:structMap>
50       <mets:div TYPE="Intellectual entity">
51         <mets:div ID="ba6865e5-7682-4cc0-8e80-0f1ccac3c028">
52           ADMID="TMD-ba6865e5-7682-4cc0-8e80-0f1ccac3c028-SCAPE_REPRESENTATION_TECHNICAL"
53           TYPE="Representation" xlink:label="">
54             <mets:fptr FILEID="ba6865e5-7682-4cc0-8e80-0f1ccac3c029" />
55           </mets:div>
56         </mets:div>
57       </mets:structMap>
58     </mets:mets>

```

Performing quality assurance on Hadoop platform



- running Jpylyzer
- comparing profile against control policy

Control Policy

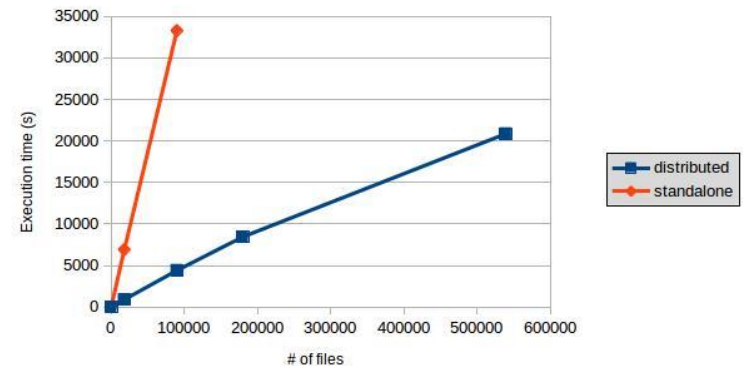
```
-<rdf:RDF xml:base="http://www.statsbiblioteket.dk/policies/">
  -<org:Organization rdf:about="danish state and university library">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#NamedIndividual"/>
    <org:identifier>The Danish State and University Library</org:identifier>
  </org:Organization>
  -<owl:NamedIndividual rdf:about="danish newspapers">
    <rdf:type rdf:resource="http://purl.org/DP/preservation-case#ContentSet"/>
  </owl:NamedIndividual>
  -<owl:NamedIndividual rdf:about="danish newspapers_scenario">
    <rdf:type rdf:resource="http://purl.org/DP/preservation-case#PreservationCase"/>
    <skos:prefLabel>Danish newspapers</skos:prefLabel>
    <!-- Colour space test -->
    <preservation-case:hasObjective rdf:resource="ColourSpaceAttribute_meth_MustBeEnumerated"/>
    <preservation-case:hasObjective rdf:resource="ColourSpaceAttribute_enumCS_MustBeGreyscale"/>
    <!-- Colour depth test -->
    <preservation-case:hasObjective rdf:resource="ColourDepthAttribute_bPCDepth_MustBe8"/>
    <preservation-case:hasUserCommunity rdf:resource="researchers"/>
    <preservation-case:hasContentSet rdf:resource="danish_newspapers"/>
  </owl:NamedIndividual>
  -<owl:NamedIndividual rdf:about="ColourSpaceAttribute_meth_MustBeEnumerated">
    <rdf:type rdf:resource="http://purl.org/DP/control-policy#FORMAT_OBJECTIVE"/>
    <skos:prefLabel>Colour space attribute 'meth' must be 'Enumerated'</skos:prefLabel>
    <control-policy:value rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Enumerated</control-policy:value>
    <control-policy:measure rdf:resource="http://purl.org/DP/quality/measures#18"/>
    <preservation-case:contentSetScope rdf:resource="danish_newspapers"/>
    <control-policy:modality rdf:resource="http://purl.org/DP/control-policy/modalities#MUST"/>
  </owl:NamedIndividual>
  -<owl:NamedIndividual rdf:about="ColourSpaceAttribute_enumCS_MustBeGreyscale">
    <rdf:type rdf:resource="http://purl.org/DP/control-policy#FORMAT_OBJECTIVE"/>
    -<skos:prefLabel>
      Colour space attribute 'enumCS' must be 'greyscale'
    </skos:prefLabel>
    <control-policy:value rdf:datatype="http://www.w3.org/2001/XMLSchema#string">greyscale</control-policy:value>
    <control-policy:measure rdf:resource="http://purl.org/DP/quality/measures#19"/>
    <preservation-case:contentSetScope rdf:resource="danish_newspapers"/>
    <control-policy:modality rdf:resource="http://purl.org/DP/control-policy/modalities#MUST"/>
  </owl:NamedIndividual>
  -<owl:NamedIndividual rdf:about="ColourDepthAttribute_bPCDepth_MustBe8">
    <rdf:type rdf:resource="http://purl.org/DP/control-policy#FORMAT_OBJECTIVE"/>
    <skos:prefLabel>Colour depth attribute 'bPCDepth' must be '8'</skos:prefLabel>
    <control-policy:value rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">8</control-policy:value>
    <control-policy:measure rdf:resource="http://purl.org/DP/quality/measures#20"/>
    <preservation-case:contentSetScope rdf:resource="danish_newspapers"/>
    <control-policy:modality rdf:resource="http://purl.org/DP/control-policy/modalities#MUST"/>
  </owl:NamedIndividual>
  -<foaf:Group rdf:about="researchers">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#NamedIndividual"/>
  </foaf:Group>
</rdf:RDF>
```

Jpylyzer Metadata

```
<mets:amdSec>
  <mets:techMD
    ID="TMD-uuid:ba6865e5-7682-4cc0-8e80-0f1ccac3c028-uuid:ba6865e5-7682-4cc0-8e80-0f1ccac3c029-jpylyzer">
      <mets:mdWrap MDTYPE="OTHER">
        <mets:xmlData>
          <jpylyzer>
            <toolInfo>
              <toolName>jpylyzer.py</toolName>
              <toolVersion>1.10.1</toolVersion>
            </toolInfo>
            <fileInfo>
              <fileName>adresseavisen1759-1795-06-13-01-0006.jp2</fileName>
              <filePath>/home/runi/development/scape-jp2-experiment/src/test/resources/sample/adresseavisen1759-1795-06-13-01-0006.jp2</filePath>
              <fileSizeInBytes>3662332</fileSizeInBytes>
              <fileLastModified>Fri Jan 24 13:36:13 2014</fileLastModified>
            </fileInfo>
            <isValidJP2>True</isValidJP2>
            <tests />
            <properties>
              <signatureBox />
              <fileTypeBox>
                <br>jp2 </br>
                <minV>0</minV>
                <cL>jp2 </cL>
              </fileTypeBox>
              <jp2HeaderBox>
                <imageHeaderBox>
                  <height>2859</height>
                  <width>2312</width>
                  <nC>1</nC>
                  <bPCSign>unsigned</bPCSign>
                  <bPCDepth>8</bPCDepth>
                  <c>jpeg2000</c>
                  <unkC>yes</unkC>
                  <iPR>no</iPR>
                </imageHeaderBox>
                <colourSpecificationBox>
                  <meth>Enumerated</meth>
                  <prec>0</prec>
                  <approx>0</approx>
                  <enumCS>greyscale</enumCS>
                </colourSpecificationBox>
                <resolutionBox>
                  <captureResolutionBox>
                    <vRcN>30000</vRcN>
                    <vRcD>254</vRcD>
                    <hRcN>30000</hRcN>
                    <hRcD>254</hRcD>
                    <vRcE>2</vRcE>
                    <hRcE>2</hRcE>
                    <vRescInPixelsPerMeter>11811.02</vRescInPixelsPerMeter>
                    <hRescInPixelsPerMeter>11811.02</hRescInPixelsPerMeter>
                    <vRescInPixelsPerInch>300.0</vRescInPixelsPerInch>
                    <hRescInPixelsPerInch>300.0</hRescInPixelsPerInch>
                  </captureResolutionBox>
                  </resolutionBox>
                </jp2HeaderBox>
              <contiguousCodestreamBox>
                <tsiz>
                  <lsiz>41</lsiz>
                  <rsiz>ISO/IEC 15444-1</rsiz>
                  <xsiz>2312</xsiz>
                  <ysiz>2859</ysiz>
                  <x0siz>0</x0siz>
                  <y0siz>0</y0siz>
                  <xTsiz>1024</xTsiz>
                  <yTsiz>1024</yTsiz>
                </tsiz>
              </contiguousCodestreamBox>
            </properties>
          </jpylyzer>
        </mets:xmlData>
      </mets:mdWrap>
    </mets:techMD>
  </mets:amdSec>
```

- Using Loader component
 - Updating DOMS objects

- Stager timings
 - work in progress
- Validation timings
- Loader timings
 - work in progress



Resources

- Newspaper Digitisation Project:
<http://en.statsbiblioteket.dk/national-library-division/newspaper-digitisation/newspaper-digitization>
- State and University Library:
<http://en.statsbiblioteket.dk/>
- Ninestars Information Technologies Ltd:
<http://ninestar.co.in/>
- Control Policy Driven Validation Experiment:
<http://wiki.opf-labs.org/display/SP/Validate+JPEG2000+Newspapers+Using+Jpylyzer>
- DOMS, fedora-based repository:
<http://www.fedora-commons.org/>
- BITMAGASIN, BitRepository:
<http://digitalbevaring.dk/det-nationale-bitmagasin/>
<https://sbforge.org/display/BITMAG/The+Bit+Repository+project>
- Apache Hadoop:
<http://hadoop.apache.org/>
- Jpylyzer:
<https://github.com/openplanets/jpylyzer>
- METS schema standard:
<http://www.loc.gov/standards/mets/>
- JPEG2000:
<http://www.jpeg.org/jpeg2000/>
- SCAPE Control Policy:
<http://wiki.opf-labs.org/display/SP/Catalogue+of+Preservation+Policy+Elements>