Note 1029051519

Herkenning Inhoud EASY.1

notitie 19 mei 2015, Henk van den Berg

Het valideren en identificeren van bestanden en het extraheren van technische metadata kan met een veelvoud aan tools. FITS (File Information Toolset) combineert een aantal van deze tools en bundelt die informatie in een xml-file per geïnspecteerd bestand. Met de C3PO command line tool kunnen we de FITS-output inlezen in MongoDB, een document database. De C3PO web-api is bedoeld als interface om de gegevens inzichtelijk te maken.  
  
In deze notitie testen we de door C3PO voorgestelde workflow over de verschillende applicaties. FITS komt daaruit als een bruikbare tool in de context van EASY; de C3PO web-api is erg beperkt en vertoont mankementen. MongoDB hanteert een vendor-specifieke query-taal.  
  
De xml-output van FITS is bruikbaar en kan als technische metadata opgenomen als inline datastream in het file-object in Fedora. De analyse die FITS uitvoert is tijdrovend en resource-intensief. Waar en wanneer kan die analyse het best worden gerund? Is Apache-Tika een alternatief voor FITS? Hoe kunnen we de gegevens over individuele bestanden verzamelen zodat standaard en ad-hoc queries vanuit verschillende vraagstellingen kunnen worden uitgevoerd?

1. C3PO aan het Werk

Eerste punt uit de opdrachtomschrijving van het project:  
*1. C3PO uitzoeken: hoe kan deze op Fedora uitgevoerd worden; zijn er aanpassingen nodig.*

## Wat is wat?

C3PO  
Clever, Crafty, Content Profiling of Objects ([c3po](https://github.com/openplanets/c3po" \t "_blank)) is a software tool prototype, which uses [FITS](http://code.google.com/p/fits" \t "_blank) generated data of a digital collection as input and generates a profile of the content set in an automatic fashion.

C3PO is een tool dat niet zelf metadata extractie uitvoert. Het gebruikt metadata geproduceerd door andere frameworks. Op dit moment zijn de xml-formats geproduceerd door FITS of Apache Tika bruikbaar als invoer voor de tool. De gegevens worden door de tool opgeslagen in de document database MongoDB in de vorm van collecties. Verder biedt dit prototype een web-gebaseerde view op de gecollecteerde metadata in de vorm van histogrammen en een view op de metadata voor een individueel bestand. Ook is er de mogelijkheid om in te zoomen op onderdelen van de verzamelde metadata (drill down) en kunnen de gegevens, al-dan-niet gefilterd, geëxporteerd worden.

**FITS**  
The File Information Tool Set (FITS) identifies, validates and extracts technical metadata for a wide range of file formats. It acts as a wrapper, invoking and managing the output from several other open source tools. Output from these tools are converted into a common format, compared to one another and consolidated into a single XML output file.

FITS is in ontwikkeling sinds 2009, de laatste release van mei 2015 hanteert nog steeds de nul in het versienummer: fits-0.8.5. FITS software kan gedownload worden als een zipfile, die de complete set van onderliggende tools en dependencies aan boord heeft. Het kan aangeroepen worden vanuit een command-line interface en biedt een Java-API. Bij gebruik van de Java-API is de output ook als JDOM-object beschikbaar.

**Apache Tika**  
The Apache Tika™ toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more.

Tika kan gebruikt worden als *parser* en als *detector*. Parsen wil in dit geval zeggen het omzetten naar tekst van de content van een bestand. *Detecting* slaat op het extraheren van metadata en bestandsherkenning. Als *detector* is Tika ook opgenomen in de reeks van FITS-tools.

MongoDB is een document database die data opslaat in BSON-structuren. BSON is de binaire variant op JSON.

**MongoDB data model**  
MongoDB stores data in the form of [BSON](http://docs.mongodb.org/manual/reference/glossary/#term-bson) documents, which are rich mappings of keys, or field names, to values. BSON supports a rich collection of types, and fields in BSON documents may hold arrays of values or embedded documents. All documents in MongoDB must be less than 16MB, which is the [BSON document size](http://docs.mongodb.org/manual/reference/limits/#BSON-Document-Size).  
  
All documents are part of a [collection](http://docs.mongodb.org/manual/reference/glossary/#term-collection), which are a logical groupings of documents in a MongoDB database. The documents in a collection share a set of indexes, and typically these documents share common fields and structure.  
  
In MongoDB the [database](http://docs.mongodb.org/manual/reference/glossary/#term-database) construct is a group of related collections. Each database has a distinct set of data files and can contain a large number of collections. A single MongoDB deployment may have many databases.

## Proefopstelling

Alle onderdelen op de door C3PO voorgestelde manier laten samenwerken, ziet er schematisch als volgt uit.



Figuur 1. Proefopstelling met C3PO. Bestanden worden door FITS geanalyseerd. Per bestand geeft dat een xml-bestand met door onderliggende tools gedetecteerde en door FITS geselecteerde gegevens. C3PO CMD-line schrijft de gegevens naar document database MongoDB; C3PO web-api geeft inzicht in de gegevens.

De proefopstelling is gerealiseerd op een virtual machine. De proefopstelling laten werken op de zandbak server, zodat ook anderen met de web-interface kunnen spelen, lukte in eerste instantie niet omdat het geheugen van de zandbak vol zit.

### Eerste observaties

De FITS command-line tool werkt probleemloos. De tool kan individuele bestanden en bestanden in een directory analyseren; een directory kan eventueel recursief doorlopen worden. De output-locatie bevat na afloop xml-bestanden met de namen van de geanalyseerde bestanden uitgebreid met .fits.xml. Bij conflicten (zelfde bestandsnaam in verschillende directories) wordt de xml-bestandsnaam uitgebreid met een nummer. Naar keuze kan de output in het fits-xml format en in een gestandaardiseerd xml-format voor bepaalde typen van bestanden. In het laatste geval wordt het gestandaardiseerde format inline opgenomen in de fits-xml. Op het moment van schrijven kent FITS gestandaardiseerde formats voor de bestandstypen audio, document, image en text. Gestandaardiseerde output voor video wordt nog niet ondersteund. Voorbeelden van FITS-output over een JPEG en een word-document zijn opgenomen in appendices 2 en 3. Het gestandaardiseerde format bevat ongeveer de zelfde gegevens als het FITS-format, values zijn soms anders geformatteerd, keys hebben een andere naam. Een gestandaardiseerd format voor een huis-tuinen-keuken Word document ontbreekt kennelijk.

De Java-API van FITS doet ietwat ouderwets aan. Source-code wordt gebouwd onder Ant. Maven integratie ontbreekt. Logging middels log4j. Gebruik van de API in de door DANS gehanteerde manier van (Java) software development en deployment heeft daarom nog wat haken en ogen, maar levert m.i. geen onoverkomelijke hindernissen. Bij een testje onder Eclipse, waarin we gebruik maakten van de al gecompileerde jar-files uit de gedownloade zip, roepen verschillende gewrapte tools dat ze hun daily-rolling.log niet kwijt kunnen; de “NLNZ metadata extractor” strooit met NullPointerExceptions die overigens het verloop van de analyse niet onderbreken.

De C3PO-tool is natuurlijk een prototype en hoort alleen al daarom niet in een professionele keten van applicaties. Dat de command-line tool is wat het is - een command-line tool – hoeft niet bezwaarlijk te zijn. Ook die kan geautomatiseerd gestart worden. Van groter belang zijn de volgende bezwaren.

1. De manier waarop C3PO de gegevens opslaat in de database is onduidelijk. Het is zeker geen een-op-een vertaling van xml-documenten naar BSON-documenten. Dit maakt uitbreiding van toepassingen gebaseerd op deze gegevensbron en ad-hoc queries problematisch, zo niet onmogelijk.
2. De MongoDB query-taal is vendor-specifiek.
3. De interface van de C3PO web-api is ontoereikend voor grote hoeveelheden bestanden. Gegevens worden ingedeeld in *collections* die middels een drop-down gekozen kunnen worden (tabblad *overview*). Het op deze manier hanteerbare aantal *collections* is daardoor beperkt tot zeg 50. De items uit een *collection*, oftewel de gegevens over individuele bestanden, zijn te benaderen vanuit een paginering van die items (tabblad *objects*). Per pagina 25 links. Bij 3 miljoen bestanden en 50 *collections* zouden er dan 2400 van die pagina’s per *collection* zijn.
4. De C3PO web-api vertoont fouten. Een nieuwe collectie wordt soms niet getoond. Bij incrementeel uitbreiden van een collectie houdt het tabblad ‘overview' het bij het oude overzicht, ook na wisseling van browser en herstarten van de server.
5. De war-file van de C3PO web-api die gedownload kan worden, werkt niet onder Tomcat. Na deployen sterft de applicatie in stilte met een 404. (In de proefopstelling werd de web-api gedeployed vanuit het play framework.)

Punten 3 en 4 van deze opsomming, dat de web-api van C3PO feitelijk onbruikbaar is, ook na eventuele reparatie van bugs, maken dat ik adviseer om niet verder te gaan met C3PO. Ook niet vanuit de optiek om al vast een out-of-the-box toy-opstelling voor bestandsherkenning en metadatering te hebben in afwachting van beter materiaal. Wat heb je aan een web-interface die het niet doet?

Hoe kan het dan wel? In het vervolg bekijken we bestands-metadatering vanuit een breder perspectief.

1. Wat nu?

Nu C3PO door de mand is gevallen kunnen we ons concentreren op de context waarbinnen we technische metadata willen bevragen, het datamodel dat daarvoor geëigend is, welke tools we waar en wanneer willen gebruiken en de manier waarop we informatie uit de verzamelde metadata kunnen halen. Hoe kunnen we op korte termijn concrete doelen halen, zonder dat dit een lange-termijn visie in de weg staat.

## Informatie-complex en Observatorium

Het archief in bredere zin is een samenstel van gebruikers, gebruik, systemen, data en metadata, archivarissen, collecties en structuren. Aspecten van dit brede samenstel kunnen we bestuderen en in kaart brengen. Dat kan vaak alleen als we gegevens uit en over het archief in context en geïntegreerd kunnen bevragen. Deze interesse kent verschillende motieven: procescontrole, management informatie, wetenschappelijke nieuwsgierigheid. Het winnen van informatie kan op ad-hoc basis of een telkens terugkerende noodzaak zijn (en daarom hopelijk geautomatiseerd). Buiten het wat en waarom van deze informatie-winning om kunnen we stellen dat er behoefte is aan een informatie-complex waar gegevens uit en over het archief worden samengebracht en dat de winning van genoemde informatie faciliteert. Uiteindelijk hebben we dan een “observatory” (quote Herbert van de Sompel) op het archief.

Het verzamelen van technische metadata over, en identificatie en validatie van bestanden in het archief hoort zeker ook tot dit informatie-complex. Enkele voorbeelden van vragen die je zou kunnen stellen met bestands-metadata als leidmotief en die je in toenemende mate alleen kunt beantwoorden in de context van, en geïntegreerd met andere gegevensbronnen.

*- Welke bestanden in het archief zijn niet in het ‘preferred format’ voor dat type bestand?*

*- Welke bestanden in het archief zijn in een verouderd format en dreigen daarom binnenkort ontoegankelijk te worden; om welke formats en typen gaat het hier?*

*- Hoe zit het met de spreiding van bestandstypen over de ‘audiences’?*

*- Is er een correlatie tussen bestandstype (text, document, image) en waar die bestanden worden gedownload (Nederland, buitenland)? In hoeverre spelen de populariteit van een dataset en de eventuele taal waarin een document is gesteld hierbij een rol? Maakt ‘audience’ een verschil in dit verband?*

De opsomming van mogelijke vragen is natuurlijk bij lange niet uitputtend – en alle mogelijke vragen willen beantwoorden hoeft niet ons doel te zijn. Merk op dat met een geïsoleerde gegevensbron (technische metadata over bestanden bijvoorbeeld) een aantal vragen adekwaat te beantwoorden is en dat bij combinatie van gegevensbronnen de groei van het aantal mogelijke vragen die beantwoord kunnen worden niet rechtevenredig maar exponentieel toeneemt.

## Een datamodel voor technische metadata

<http://projects.iq.harvard.edu/fits/fits-xml>

<http://en.wikipedia.org/wiki/Extensible_Metadata_Platform>

Nog meer libraries voor metadata extractie:

<https://github.com/drewnoakes/metadata-extractor> (image files)

## Schetsen voor een Systeemarchitectuur

- Processing time for ~1 Gb of data (folder ‘original’ of dataset ...) 8 minutes

Appendix 1. – Opzetten van een VM

De applicaties MongoDB, C3PO command line tool, C3PO web api en FITS werden getest op een Virtual Machine. Wat volgt is het relaas van de installatie van deze applicaties, zodat installatie reproduceerbaar is. Het is zeker niet bedoeld als de definitieve installatiegids.

Vagrant VM

We started with easy-dtap, downloaded 2015-05-11. See for instructions  
<http://develop.dans.knaw.nl/easy/2015/04/15/easy-dtap-gebruik-alleen-deasy.html>

Before vagrant up, added port 9000 to the Vagrantfile:

deasy.vm.network "forwarded\_port", guest: 9000, host: 19000

Changed to root user.

See also c3po usage guide: <https://github.com/peshkira/c3po/wiki/Usage-Guide>

JDK

The play framework needs javac in order to compile source code of the c3po-web-api. Downloaded jdk 7 from the Oracle web site in the shared folder /vagrant.

# rpm -ivh jdk-7u79-linux-x64.rpm

# alternatives --install /usr/bin/java java /usr/java/jdk1.7.0\_79/bin/java 2

# alternatives --config java

# java -version

java version "1.7.0\_79"

Java(TM) SE Runtime Environment (build 1.7.0\_79-b15)

Java HotSpot(TM) 64-Bit Server VM (build 24.79-b02, mixed mode)

# javac -version

javac 1.7.0\_79

MongoDB

C3po command line tool and web-api need a running mongodb.  
See: <http://docs.mongodb.org/manual/tutorial/install-mongodb-on-red-hat/>

# vi /etc/yum.repos.d/mongodb-org-3.0.repo

with text:

[mongodb-org-3.0]

name=MongoDB Repository

baseurl=http://repo.mongodb.org/yum/redhat/$releasever/mongodb-org/3.0/x86\_64/

gpgcheck=0

enabled=1

# yum install mongodb-org  
 Installed:

mongodb-org.x86\_64 0:3.0.2-1.el6

# service mongod start

FITS

See: <http://projects.iq.harvard.edu/fits>

mkdir /opt/fits  
cd /opt/fits/  
wget <http://projects.iq.harvard.edu/files/fits/files/fits-0.8.5.zip>

# unzip fits-0.8.5.zip

# cd fits-0.8.5

# chmod =x fits.sh

C3PO command line tool

mkdir /opt/c3po  
cd /opt/c3po/  
wget <https://bintray.com/artifact/download/peshkira/c3po/cli/c3po-cmd-0.4.0.jar>

Play Framework

# mkdir /opt/play

# cd /opt/play/

# wget <http://downloads.typesafe.com/releases/play-2.0.4.zip>

# unzip play-2.0.4.zip

# export PATH=$PATH:/opt/play/play-2.0.4

C3PO web-api

Forked the <https://github.com/peshkira/c3po> and cloned it in shared folder /vagrant.

# cd /vagrant/c3po/c3po-webapi/

# play clean compile stage

...

[warn] Note: /vagrant/c3po/c3po-webapi/app/controllers/FilterController.java uses or overrides a deprecated API.

[warn] Note: Recompile with -Xlint:deprecation for details.

[warn] Note: Some input files use unchecked or unsafe operations.

[warn] Note: Recompile with -Xlint:unchecked for details.

[success] Total time: 33 s, completed May 13, 2015 10:57:19 AM

[info] Packaging /vagrant/c3po/c3po-webapi/target/scala-2.9.1/c3po\_2.9.1-0.3.0.jar ...

[info] Done packaging.

[info]

[info] Your application is ready to be run in place: target/start

[info]

[success] Total time: 4 s, completed May 13, 2015 10:57:23 AM

All together now

Important: Do not start the C3PO web-api before the database is initialized with the C3PO command line tool.

Run FITS over the RELEASE.txt:

# cd /opt/fits/fits-0.8.5  
# mkdir /tmp/fits

# ./fits.sh -i RELEASE.txt -o /tmp/fits/fits\_out.txt

Gather the output with c3po cmd line:

# cd /opt/c3po

# java -jar c3po-cmd-0.4.0.jar gather -c test123 -i /tmp/fits/

Start the web api:

# cd /vagrant/c3po/c3po-webapi/

# play

[info] Loading project definition from /vagrant/c3po/c3po-webapi/project

[info] Set current project to c3po (in build file:/vagrant/c3po/c3po-webapi/)

\_ \_

\_ \_\_ | | \_\_ \_ \_ \_| |

| '\_ \| |/ \_' | || |\_|

| \_\_/|\_|\\_\_\_\_|\\_\_ (\_)

|\_| |\_\_/

play! 2.0.4, http://www.playframework.org

> Type "help play" or "license" for more information.

> Type "exit" or use Ctrl+D to leave this console.

[c3po] $ run

--- (Running the application from SBT, auto-reloading is enabled) ---

[info] play - Listening for HTTP on port 9000...

(Server started, use Ctrl+D to stop and go back to the console...)

--> Web api running under <http://localhost:19000/>

Appendix 2. – JPEG voorbeeld

Input: jpeg, foto genomen met een GPS-camera.

Output in FITS-xml

<?xml version="1.0" encoding="UTF-8"?>  
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits\_output" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits\_output http://hul.harvard.edu/ois/xml/xsd/fits/fits\_output.xsd" version="0.8.5" timestamp="5/19/15 10:08 AM">  
 <identification>  
 <identity format="JPEG File Interchange Format" mimetype="image/jpeg" toolname="FITS" toolversion="0.8.5">  
 <tool toolname="Jhove" toolversion="1.5" />  
 <tool toolname="file utility" toolversion="5.04" />  
 <tool toolname="Exiftool" toolversion="9.13" />  
 <tool toolname="Droid" toolversion="6.1.3" />  
 <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />  
 <version toolname="Jhove" toolversion="1.5">1.01</version>  
 <externalIdentifier toolname="Droid" toolversion="6.1.3" type="puid">fmt/43</externalIdentifier>  
 </identity>  
 </identification>  
 <fileinfo>  
 <size toolname="Jhove" toolversion="1.5">730945</size>  
 <creatingApplicationName toolname="Jhove" toolversion="1.5">Montana 650</creatingApplicationName>  
 <lastmodified toolname="Exiftool" toolversion="9.13" status="CONFLICT">2015:05:19 10:15:04+02:00</lastmodified>  
 <lastmodified toolname="Tika" toolversion="1.3" status="CONFLICT">2013-10-04T14:53:31</lastmodified>  
 <created toolname="NLNZ Metadata Extractor" toolversion="3.4GA" status="SINGLE\_RESULT">2013:10:04 14:53:31</created>  
 <filepath toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">/vagrant/files/test1/DSC00323.jpg</filepath>  
 <filename toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">DSC00323.jpg</filename>  
 <md5checksum toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">9592dcfeca297b110cbf64686edeaba9</md5checksum>  
 <fslastmodified toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">1432023304000</fslastmodified>  
 </fileinfo>  
 <filestatus>  
 <well-formed toolname="Jhove" toolversion="1.5" status="SINGLE\_RESULT">true</well-formed>  
 <valid toolname="Jhove" toolversion="1.5" status="SINGLE\_RESULT">true</valid>  
 </filestatus>  
 <metadata>  
 <image>  
 <byteOrder toolname="Jhove" toolversion="1.5" status="SINGLE\_RESULT">big endian</byteOrder>  
 <compressionScheme toolname="Jhove" toolversion="1.5" status="SINGLE\_RESULT">JPEG (old-style)</compressionScheme>  
 <imageWidth toolname="Jhove" toolversion="1.5">1200</imageWidth>  
 <imageHeight toolname="Jhove" toolversion="1.5">1600</imageHeight>  
 <colorSpace toolname="Jhove" toolversion="1.5" status="SINGLE\_RESULT">YCbCr</colorSpace>  
 <iccProfileName toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">sRGB IEC61966-2.1</iccProfileName>  
 <YCbCrSubSampling toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">2 1</YCbCrSubSampling>  
 <orientation toolname="Exiftool" toolversion="9.13">normal\*</orientation>  
 <samplingFrequencyUnit toolname="Jhove" toolversion="1.5" status="SINGLE\_RESULT">in.</samplingFrequencyUnit>  
 <xSamplingFrequency toolname="Jhove" toolversion="1.5">72</xSamplingFrequency>  
 <ySamplingFrequency toolname="Jhove" toolversion="1.5">72</ySamplingFrequency>  
 <bitsPerSample toolname="Jhove" toolversion="1.5">8 8 8</bitsPerSample>  
 <samplesPerPixel toolname="Jhove" toolversion="1.5" status="SINGLE\_RESULT">3</samplesPerPixel>  
 <lightSource toolname="Jhove" toolversion="1.5">unknown</lightSource>  
 <iccProfileVersion toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">2.1.0</iccProfileVersion>  
 <scannerManufacturer toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">Garmin</scannerManufacturer>  
 <scannerModelName toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">Montana 650</scannerModelName>  
 <exifVersion toolname="Exiftool" toolversion="9.13">0220</exifVersion>  
 <gpsVersionID toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">2.2.0.0</gpsVersionID>  
 <gpsLatitudeRef toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">N</gpsLatitudeRef>  
 <gpsLatitude toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">52 deg 14' 56.75" N</gpsLatitude>  
 <gpsLongitudeRef toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">E</gpsLongitudeRef>  
 <gpsLongitude toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">5 deg 8' 18.90" E</gpsLongitude>  
 <gpsAltitudeRef toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">Sea level</gpsAltitudeRef>  
 <gpsAltitude toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">42.7 m Above Sea Level</gpsAltitude>  
 <gpsImgDirectionRef toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">T</gpsImgDirectionRef>  
 <gpsImgDirection toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">160.041769</gpsImgDirection>  
 <gpsMapDatum toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">WGS-84</gpsMapDatum>  
 </image>  
 </metadata>  
 <statistics fitsExecutionTime="4325">  
 <tool toolname="OIS Audio Information" toolversion="0.1" status="did not run" />  
 <tool toolname="ADL Tool" toolversion="0.1" status="did not run" />  
 <tool toolname="Jhove" toolversion="1.5" executionTime="3894" />  
 <tool toolname="file utility" toolversion="5.04" executionTime="3447" />  
 <tool toolname="Exiftool" toolversion="9.13" executionTime="3845" />  
 <tool toolname="Droid" toolversion="6.1.3" executionTime="415" />  
 <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" executionTime="3046" />  
 <tool toolname="OIS File Information" toolversion="0.2" executionTime="404" />  
 <tool toolname="OIS XML Metadata" toolversion="0.2" status="did not run" />  
 <tool toolname="ffident" toolversion="0.2" executionTime="1514" />  
 <tool toolname="Tika" toolversion="1.3" executionTime="3142" />  
 </statistics>  
</fits>

Standard xml (in dit geval MIX, NISO Metadata for Images in XML) is opgenomen in FITS-xml. (volgende pagina) Standaard xml geeft in feite de zelfde informatie in een ander format.

<?xml version="1.0" encoding="UTF-8"?>  
...  
 <standard>  
 <mix:mix xmlns:mix="http://www.loc.gov/mix/v20">  
 <mix:BasicDigitalObjectInformation>  
 <mix:byteOrder>big endian</mix:byteOrder>  
 <mix:Compression>  
 <mix:compressionScheme>JPEG (old-style)</mix:compressionScheme>  
 </mix:Compression>  
 </mix:BasicDigitalObjectInformation>  
 <mix:BasicImageInformation>  
 <mix:BasicImageCharacteristics>  
 <mix:imageWidth>1200</mix:imageWidth>  
 <mix:imageHeight>1600</mix:imageHeight>  
 <mix:PhotometricInterpretation>  
 <mix:colorSpace>YCbCr</mix:colorSpace>  
 <mix:ColorProfile>  
 <mix:IccProfile>  
 <mix:iccProfileName>sRGB IEC61966-2.1</mix:iccProfileName>  
 <mix:iccProfileVersion>2.1.0</mix:iccProfileVersion>  
 </mix:IccProfile>  
 </mix:ColorProfile>  
 <mix:YCbCr>  
 <mix:YCbCrSubSampling>  
 <mix:yCbCrSubsampleHoriz>2</mix:yCbCrSubsampleHoriz>  
 <mix:yCbCrSubsampleVert>1</mix:yCbCrSubsampleVert>  
 </mix:YCbCrSubSampling>  
 </mix:YCbCr>  
 </mix:PhotometricInterpretation>  
 </mix:BasicImageCharacteristics>  
 </mix:BasicImageInformation>  
 <mix:ImageCaptureMetadata>  
 <mix:GeneralCaptureInformation>  
 <mix:dateTimeCreated>2013-10-04T12:53:31.0Z</mix:dateTimeCreated>  
 </mix:GeneralCaptureInformation>  
 <mix:ScannerCapture>  
 <mix:scannerManufacturer>Garmin</mix:scannerManufacturer>  
 <mix:ScannerModel>  
 <mix:scannerModelName>Montana 650</mix:scannerModelName>  
 </mix:ScannerModel>  
 <mix:ScanningSystemSoftware />  
 </mix:ScannerCapture>  
 <mix:DigitalCameraCapture>  
 <mix:DigitalCameraModel />  
 <mix:CameraCaptureSettings>  
 <mix:ImageData>  
 <mix:exifVersion>0220</mix:exifVersion>  
 <mix:lightSource>unknown</mix:lightSource>  
 </mix:ImageData>  
 <mix:GPSData>  
 <mix:gpsVersionID>2.2.0.0</mix:gpsVersionID>  
 <mix:gpsLatitudeRef>N</mix:gpsLatitudeRef>  
 <mix:GPSLatitude>  
 <mix:degrees>  
 <mix:numerator>52</mix:numerator>  
 <mix:denominator>1</mix:denominator>  
 </mix:degrees>  
 <mix:minutes>  
 <mix:numerator>14</mix:numerator>  
 <mix:denominator>1</mix:denominator>  
 </mix:minutes>  
 <mix:seconds>  
 <mix:numerator>5675</mix:numerator>  
 <mix:denominator>100</mix:denominator>  
 </mix:seconds>  
 </mix:GPSLatitude>  
 <mix:gpsLongitudeRef>E</mix:gpsLongitudeRef>  
 <mix:GPSLongitude>  
 <mix:degrees>  
 <mix:numerator>5</mix:numerator>  
 <mix:denominator>1</mix:denominator>  
 </mix:degrees>  
 <mix:minutes>  
 <mix:numerator>8</mix:numerator>  
 <mix:denominator>1</mix:denominator>  
 </mix:minutes>  
 <mix:seconds>  
 <mix:numerator>1889</mix:numerator>  
 <mix:denominator>100</mix:denominator>  
 </mix:seconds>  
 </mix:GPSLongitude>  
 <mix:gpsAltitudeRef>Sea level</mix:gpsAltitudeRef>  
 <mix:gpsImgDirectionRef>T</mix:gpsImgDirectionRef>  
 <mix:gpsImgDirection>  
 <mix:numerator>16004</mix:numerator>  
 <mix:denominator>100</mix:denominator>  
 </mix:gpsImgDirection>  
 <mix:gpsMapDatum>WGS-84</mix:gpsMapDatum>  
 </mix:GPSData>  
 </mix:CameraCaptureSettings>  
 </mix:DigitalCameraCapture>  
 <mix:orientation>normal\*</mix:orientation>  
 </mix:ImageCaptureMetadata>  
 <mix:ImageAssessmentMetadata>  
 <mix:SpatialMetrics>  
 <mix:samplingFrequencyUnit>in.</mix:samplingFrequencyUnit>  
 <mix:xSamplingFrequency>  
 <mix:numerator>72</mix:numerator>  
 <mix:denominator>1</mix:denominator>  
 </mix:xSamplingFrequency>  
 <mix:ySamplingFrequency>  
 <mix:numerator>72</mix:numerator>  
 <mix:denominator>1</mix:denominator>  
 </mix:ySamplingFrequency>  
 </mix:SpatialMetrics>  
 <mix:ImageColorEncoding>  
 <mix:BitsPerSample>  
 <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue>  
 <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue>  
 <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue>  
 <mix:bitsPerSampleUnit>integer</mix:bitsPerSampleUnit>  
 </mix:BitsPerSample>  
 <mix:samplesPerPixel>3</mix:samplesPerPixel>  
 </mix:ImageColorEncoding>  
 </mix:ImageAssessmentMetadata>  
 </mix:mix>  
 </standard>  
...  
</fits>

Output van Apache-Tika, output voor de keys Blue, Green en Red TRC weggelaten (...).

Application Record Version: 2

Blue Colorant: (0.1430664, 0.06060791, 0.71409607)

Blue TRC: 0.0, 0.0000763,...

CMM Type: Lino

Caption Digest: -54 -4 65 7 116 -16 -14 -122 -40 96 50 -76 -35 -10 -58 -16

Class: Display Device

Coded Character Set: UTF-8

Color Space: sRGB

Color space: RGB

Component 1: Y component: Quantization table 0, Sampling factors 1 horiz/2 vert

Component 2: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert

Component 3: Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert

Compression Type: Baseline

Content-Length: 730945

Content-Type: image/jpeg

Copyright: Copyright (c) 1998 Hewlett-Packard Company

Creation-Date: 2013-10-04T14:53:31

Data Precision: 8 bits

Date Created: Fri Oct 04 00:00:00 CEST 2013

Date/Time: 2013:10:04 14:53:31

Date/Time Original: 2013:10:04 14:53:31

Device Mfg Description: IEC http://www.iec.ch

Device Model Description: IEC 61966-2.1 Default RGB colour space - sRGB

Device manufacturer: IEC

Device model: sRGB

Exif Image Height: 1600 pixels

Exif Image Width: 1200 pixels

Exif Version: 2.20

File Modified Date: Tue May 19 10:15:04 CEST 2015

File Name: DSC00323.jpg

File Size: 730945 bytes

FlashPix Version: 1.00

GPS Altitude: 42 metres

GPS Altitude Ref: Sea level

GPS Img Direction: 160.04 degrees

GPS Img Direction Ref: True direction

GPS Latitude: 52° 14' 56.75"

GPS Latitude Ref: N

GPS Longitude: 5° 8' 18.9"

GPS Longitude Ref: E

GPS Map Datum: WGS-84

GPS Version ID: 2.200

Green Colorant: (0.3851471, 0.71687317, 0.097076416)

Green TRC: 0.0, 0.0000763, ...

Image Height: 1600 pixels

Image Width: 1200 pixels

Last-Modified: 2013-10-04T14:53:31

Last-Save-Date: 2013-10-04T14:53:31

Luminance: (76.03647, 80.0, 87.12462)

Make: Garmin

Measurement: 1931 2° Observer, Backing (0.0, 0.0, 0.0), Geometry Unknown, Flare 1%, Illuminant D65

Media Black Point: (0.0, 0.0, 0.0)

Media White Point: (0.9504547, 1.0, 1.0890503)

Model: Montana 650

Number of Components: 3

Orientation: Top, left side (Horizontal / normal)

Primary Platform: Microsoft Corporation

Profile Connection Space: XYZ

Profile Date/Time: Mon Mar 09 07:49:00 CET 1998

Profile Description: sRGB IEC61966-2.1

Profile Size: 3144

Red Colorant: (0.43606567, 0.2224884, 0.013916016)

Red TRC: 0.0, 0.0000763, ...

Resolution Units: inch

Signature: acsp

Tag Count: 17

Technology: CRT

Time Created: 14:53:31

Viewing Conditions: view(0x76696577): 36 bytes

Viewing Conditions Description: Reference Viewing Condition in IEC61966-2.1

X Resolution: 72 dots

X-Parsed-By: org.apache.tika.parser.DefaultParser

X-Parsed-By: org.apache.tika.parser.jpeg.JpegParser

XMP Value Count: 2

XYZ values: 0.9642029 1.0 0.8249054

Y Resolution: 72 dots

date: 2013-10-04T14:53:31

dcterms:created: 2013-10-04T14:53:31

dcterms:modified: 2013-10-04T14:53:31

exif:DateTimeOriginal: 2013-10-04T14:53:31

geo:lat: 52.249097

geo:long: 5.138583

meta:creation-date: 2013-10-04T14:53:31

meta:save-date: 2013-10-04T14:53:31

modified: 2013-10-04T14:53:31

resourceName: DSC00323.jpg

tiff:BitsPerSample: 8

tiff:ImageLength: 1600

tiff:ImageWidth: 1200

tiff:Make: Garmin

tiff:Model: Montana 650

tiff:Orientation: 1

Appendix 3. – FITS-output, docx voorbeeld

Input: docx, word document gemaakt met MS Word for Mac 2011, versie 14.4.9.

Output in FITS-xml

<?xml version="1.0" encoding="UTF-8"?>  
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits\_output" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits\_output http://hul.harvard.edu/ois/xml/xsd/fits/fits\_output.xsd" version="0.8.5" timestamp="5/19/15 10:08 AM">  
 <identification status="CONFLICT">  
 <identity format="OpenDocument Text" mimetype="application/vnd.oasis.opendocument.text" toolname="FITS" toolversion="0.8.5">  
 <tool toolname="file utility" toolversion="5.04" />  
 <tool toolname="Exiftool" toolversion="9.13" />  
 <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />  
 <version toolname="file utility" toolversion="5.04">2.0</version>  
 </identity>  
 <identity format="Office Open XML Document" mimetype="application/vnd.openxmlformats-officedocument.wordprocessingml.document" toolname="FITS" toolversion="0.8.5">  
 <tool toolname="Tika" toolversion="1.3" />  
 </identity>  
 </identification>  
 <fileinfo>  
 <lastmodified toolname="Exiftool" toolversion="9.13" status="SINGLE\_RESULT">2015:05:19 10:17:59+02:00</lastmodified>  
 <filepath toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">/vagrant/files/test1/Herkenning Inhoud EASY.docx</filepath>  
 <filename toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">Herkenning Inhoud EASY.docx</filename>  
 <size toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">81319</size>  
 <md5checksum toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">0a7deb87668d9637b52607b12cd41e8d</md5checksum>  
 <fslastmodified toolname="OIS File Information" toolversion="0.2" status="SINGLE\_RESULT">1432023479000</fslastmodified>  
 </fileinfo>  
 <filestatus />  
 <metadata>  
 <document />  
 </metadata>  
 <statistics fitsExecutionTime="5755">  
 <tool toolname="OIS Audio Information" toolversion="0.1" status="did not run" />  
 <tool toolname="ADL Tool" toolversion="0.1" status="did not run" />  
 <tool toolname="Jhove" toolversion="1.5" executionTime="5727" />  
 <tool toolname="file utility" toolversion="5.04" executionTime="316" />  
 <tool toolname="Exiftool" toolversion="9.13" executionTime="497" />  
 <tool toolname="Droid" toolversion="6.1.3" executionTime="43" />  
 <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" executionTime="229" />  
 <tool toolname="OIS File Information" toolversion="0.2" executionTime="14" />  
 <tool toolname="OIS XML Metadata" toolversion="0.2" status="did not run" />  
 <tool toolname="ffident" toolversion="0.2" executionTime="30" />  
 <tool toolname="Tika" toolversion="1.3" executionTime="2519" />  
 </statistics>  
</fits>

De standaard xml voor .docx ontbreekt. FITS geeft in beide modi de zelfde output.

Output van Apache-Tika:

Application-Name: Microsoft Macintosh Word

Application-Version: 14.0000

Author: henk van den berg

Character Count: 8325

Character-Count-With-Spaces: 9766

Content-Length: 81319

Content-Type: application/vnd.openxmlformats-officedocument.wordprocessingml.document

Creation-Date: 2015-05-12T08:49:00Z

Last-Author: henk van den berg

Last-Modified: 2015-05-19T08:17:00Z

Last-Save-Date: 2015-05-19T08:17:00Z

Line-Count: 69

Page-Count: 8

Paragraph-Count: 19

Revision-Number: 1

Template: report-1.1.dotx

Total-Time: 99

Word-Count: 1460

X-Parsed-By: org.apache.tika.parser.DefaultParser

X-Parsed-By: org.apache.tika.parser.microsoft.ooxml.OOXMLParser

cp:revision: 1

creator: henk van den berg

date: 2015-05-19T08:17:00Z

dc:creator: henk van den berg

dc:publisher: DANS-KNAW

dcterms:created: 2015-05-12T08:49:00Z

dcterms:modified: 2015-05-19T08:17:00Z

extended-properties:AppVersion: 14.0000

extended-properties:Application: Microsoft Macintosh Word

extended-properties:Company: DANS-KNAW

extended-properties:Template: report-1.1.dotx

extended-properties:TotalTime: 99

meta:author: henk van den berg

meta:character-count: 8325

meta:character-count-with-spaces: 9766

meta:creation-date: 2015-05-12T08:49:00Z

meta:last-author: henk van den berg

meta:line-count: 69

meta:page-count: 8

meta:paragraph-count: 19

meta:save-date: 2015-05-19T08:17:00Z

meta:word-count: 1460

modified: 2015-05-19T08:17:00Z

publisher: DANS-KNAW

resourceName: Herkenning Inhoud EASY.docx

xmpTPg:NPages: 8

Glossary

TRC

*Tone Reproduction Curve.   
In the theory of* [*photography*](http://en.wikipedia.org/wiki/Photography)*,* ***tone reproduction*** *is the mapping of scene* [*luminance*](http://en.wikipedia.org/wiki/Luminance) *and* [*color*](http://en.wikipedia.org/wiki/Color) *to print* [*reflectance*](http://en.wikipedia.org/wiki/Reflectance) *or* [*display*](http://en.wikipedia.org/wiki/Display_device)[*luminance*](http://en.wikipedia.org/wiki/Luminance)*,*[*[1]*](http://en.wikipedia.org/wiki/Tone_reproduction#cite_note-1) *with the aim of subjectively "properly" reproducing* [*brightness*](http://en.wikipedia.org/wiki/Brightness) *and "brightness differences."*[*http://en.wikipedia.org/wiki/Tone\_reproduction*](http://en.wikipedia.org/wiki/Tone_reproduction)