Stand: 07.07.2014
SH

Kernset technischer Metadaten für die Langzeitarchivierung digitaler Objekte

Version 1.1

Dokumenthistorie

Datum	Version	Bearbeitet von	Änderung
23.08.2012	0.1	Stefan Hein	Dokumenterstellung
29.01.2013	0.2	Stefan Hein	Überarbeitung / Aktualisierung
07.07.2014	1.0	Stefan Hein	Finalisierung
07.07.2014	1.1	Stefan Hein	Veröffentlichung

Hintergrund

Mit der Produktivnahme des DNB-Import-Services V2.1.0 im Rahmen des DNB-internen Projekts *DP4lib engaged* erfolgt die Generierung technischer Metadaten zu jedem Dateiobjekt als Bestandteil der Verarbeitungskette innerhalb der Import-Verarbeitung. Die gewonnenen Metadaten werden zunächst als XML-Zeichenkette im DNB-Repository als Dateieigenschaft hinterlegt und in einem asynchronen zweiten Schritt durch den Workflow „LZA-Anbindung“ weiterverarbeitet. Die LZA-Anbindung übernimmt hierbei das Erstellen und Übertragen von UOF-SIPs an das Langzeitarchiv DIAS. Beim Erstellen von UOF-SIPs werden diese im Rahmen der Import-Routinen gewonnenen Metadaten in einer METS-Datei an der entsprechenden [1]Stelle dateibezogen vermerkt.

Für die Durchführung von Preservation Actions (z. B. Migration) sind technische Metadaten für die Identifikation der zu migrierenden Objekte mithilfe von charakteristischen Eigenschaften wie Dateiformat, Formatversion, Dateigröße etc. und für Qualitätssicherungsmaßnahmen nach einer Migration notwendig. Bei Letzterem unterstützten technische Metadaten beim Vergleich von signifikanten Eigenschaften von Original und Migrationsobjekt.

Ausgangspunkt zur Identifikation der zu migrierenden Objekte ist die gezielte Anfrage an das Data-Management. Das Data-Management ist innerhalb von DIAS als DB2-Datenbank in Verbindung mit einer ergänzenden XML-Datenbank implementiert, die auch die gezielt Anfragen auf technischen Metadaten mithilfe von XPATH-Ausdrücken erlaubt.

Tools zur automatisierten Generierung technischer Metadaten

Für die Generierung kommen folgende Tools zum Einsatz:

- *didigo* V 1.0 (diagnose digital objects): Steuerung des FITS-Tools, Weiterverarbeitung des FITS-Outputs und Ableiten eines Ingest-Levels
- *FITS[2] (File Information Tools Set) V. 0.6.1*

Das FITS-Tool generiert selbst keine Metadaten, sondern bringt eine Reihe etablierten Metadatentools zum Einsatz. Dazu zählen aktuell:

- AudioInfo
- ADLTool
- Jhove 1
- FileUtility
- Exiftool
- Droid
- MetadataExtractor
- FileInfo
- FFIdent

DNB hat den FileAnalyzer - eine Eigenentwicklung - für die Unterstützung des ePub-Formats ergänzt.

[1] Vgl. [UOF], [LMER] und [UOF]
[2] Vgl. [FITS]

Nach Aufruf aller Metadaten-Tools normalisiert FITS die verschiedenen Outputs zu einem vereinheitlichten FITS-Output. Die Normalisierung findet durch eine XSLT-Transformation statt, basiert also auf der Verarbeitung von XML-kodierten Output-Informationen. Die Definition von Struktur und verwendeten Elementen des damit erstellten FITS-Outputs bildet den Inhalt des *Kernsets technischer Metadaten zur Langzeitarchivierung* und wird im folgenden Kapital näher beschrieben.

Vom FITS-Output zum Kernset technischer Metadaten V. 1.0

- XML-Schema des FITS-Outputs vgl. [FITS-Schema].

Der FITS-Output gliedert sich wie folgt:

`FITS`

`identification`
Informationen zum Dateiformat und Formatversion
- PUID
- MimeType
- Formatname

`fileinfo`
Formatunabhängige Metadaten z. B.
- Dateiname
- Dateigröße

`filestatus`
Informationen zur Formatvalidität
- Well-formed
- Valid
- Message (Fehlermeldungen)
- Ingest-level

`Metadata`
Technische Metadaten für die Objekttypen:
- Image
- Text
- Document
- Audio
- Video
- Exceutable

`toolOutput`
Optional kann der Original-Output der eingesetzten Tools mit übergeben werden. In der DNB-Konfiguration wird diese Option deaktiviert.

Normalisierung

Für jedes in FITS eingesetzte Tool ist die Angabe einer XSLT-Transformation notwendig, die Elementnamen und auch Werte in die von FITS-Output definierte Form überführt. Zum Teil wurden hierfür für Objekttypen und auch für Dateien separate XSLT-Dateien erstellt. Den Schritt dieser Normalisierung (nach FITS die sog. Konsolidierung) wird von FITS selbst durchgeführt.

Elemente des Abschnitts `metadata`

Das Default-FITS-Output-Schema [FITS-Schema] schränkt die für die versch. Objekttypen verwendeten Elemente **nicht** ein.

Das FITS-Schema [FITS-techElemtents] definiert hingegen ein festes Elementset für jedes der genannten Objekttypen innerhalb des Abschnitts `metadata` und ist Ausgangspunkt der Anpassungen der DNB zum DNB-*Kernset technischer Metadaten für die Langzeitarchivierung*.

DNB-Anpassungen

Ausgehend vom strikteren FITS-Schema [FITS-techElements] wurden folgende Ergänzungen und der damit notwendigen Mappings durchgeführt. Ergebnis dieser Erweiterung ist ein DNB-eigenes FITS-Schema [FITS-DNB-Schema] und in Folge dessen angepasste XSLT-Dateien.

Folgende Vorüberlegungen wurden hierbei getroffen, die zum einen die an der DNB zu berücksichtigenden Dateiformate als auch die zu ergänzenden Elemente bzw. Anpassungen des bereits vorhandenen Mappings berühren.

- Basis sollte zur Ausgabe `fits_output_tech_elements.xsd` sein, denn dort finden sich bereits die wichtigsten Elemente. Der Abschnitt `<toolOutput>` würde komplett entfallen.
- Die Kategorisierung in `Image`, `Text`, `Document` und `Audio` ist sinnvoll. Hinzukommen sollte `Video` und `Executable` für Programmdateien, wobei letzteres später noch zu ergänzen wäre und dabei insbesondere Emulationen und die Ergebnisse aus dem Projekt KEEP[3] zu berücksichtigen sind.
- `Document` sollte für Office-Formate, PDFs und E-Books genutzt werden. Obwohl HTML auch für diese Kategorie sinnvoll sein könnte, soll es als Text behandelt werden. XML kann abhängig von der Nutzung als spezifisches Format ein `Document` sein, aber ansonsten ist es Text.
- Ziel sollte sein, die Felder in den Kategorien für jedes zugeordnete Dateiformat soweit wie möglich zu füllen. Obwohl sich die Elemente aus pragmatischen Gründen an den real existierenden Ausgaben der Tools für bestimmte Formate orientieren, ist eine Vereinheitlichung etwa durch Summenbildungen anzustreben. Allerdings muss Raum für Elemente sein, die nur bei spezifischen Formaten der Kategorie sinnvoll sind, etwa `isPdfA`. Es gibt keine Pflichtelemente in den Metadaten. Spätere Erweiterungen sollten jederzeit möglich sein, allerdings sollten vorhandene Elemente nicht mehr geändert werden.

[3] http://www.dnb.de/DE/Wir/Projekte/Abgeschlossen/keep.html

neue Elemente nach Objekttyp, Dateiformat und Tool

PDF	alle Elemente von documentMetadata aus fits_output_tech_elements.xsd	isPdfA (konform zu PDF/A-ISO-Norm)	pdfAVersion (gebildet aus den XMP-Werten im JHOVE-Output pdfaid:part und pdfaid:conformance)	graphicsCount	
ePub	alle Elemente von documentMetadata aus fits_output_tech_elements.xsd	author	title	language	isRightsManaged
XML	alle Elemente von textMetadata aus fits_output_tech_elements.xsd				
HTML	alle Elemente von textMetadata aus fits_output_tech_elements.xsd	paragraphCount(NLNZ-PARAGRAPHS)	usesStyleSheets (NLNZ)	characterCount (NLNZ-CHARACTERS)	wordCount (NLNZ-WORDS)
PS	alle Elemente von textMetadata aus fits_output_tech_elements.xsd	creatingApplication	PageCount	charSet	
DOC	alle Elemente von documentMetadata aus fits_output_tech_elements.xsd (soweit anwendbar bzw. aus Output zu ersehen)	paragraphCount (Exif)	LineCount (Exif)		

TXT	alle Elemente von textMetadata aus fits_output_tech_elements.xsd												
Bildformate													
JPEG	alle Elemente von imageMetadata aus fits_output_tech_elements.xsd												
TIFF	alle Elemente von imageMetadata aus fits_output_tech_elements.xsd												
PNG	alle Elemente von imageMetadata aus fits_output_tech_elements.xsd	bitsPerSample (EXIF < BitDepth)	colorSpace (EXIF colorType)										
GIF	alle Elemente von imageMetadata aus fits_output_tech_elements.xsd												
Audio													
WAV	alle Elemente von audioMetadata aus fits_output_tech_elements.xsd	compressionScheme	AverageBytesPerSecond										
MP3	alle Elemente von audioMetadata aus fits_output_tech_elements.xsd	compressionScheme	ChannelMode										
Video													
MPEG	duration	bitRate	frameRate	bitDepth	sampleRate	channels	imageWidth	imageHeight	xSamplingFrequency	ySamplingFrequency	dataFormatType	blockSizeMin	blockSizeMax

Ausgehend von FITS Version 0.6.1 und dem Schema [FITS-techElements]wurden folgende Ergänzungen durchgeführt (die Ergänzungen zu vorhandenen Mappings sind in der XSLT-Mapping-Datei durch das Kommentar `<!-- added by DNB -->` markiert):

`fileStatus` (neu für ePub)

Format	Elementname	Beschreibung	XSLT-Mapping-Datei
ePub	well-formed	wohlgeformt	[FITS-HOME]/xml/fileAnalyzer/snippets/filestatus.xsl
	valid	valide	[FITS-HOME]/xml/fileAnalyzer/snippets/filestatus.xsl
	message	Fehlermeldung	[FITS-HOME]/xml/fileAnalyzer/snippets/filestatus.xsl
	countInvalid	Zahl der invaliden Bestandteile	[FITS-HOME]/xml/fileAnalyzer/snippets/filestatus.xsl

`fileInfo`

Format	Elementname	Beschreibung	XSLT-Mapping-Datei
PS	creatingApplication	Erzeuger-anwendung	[FITS-HOME/xml/exiftool/exiftool_common_to_fits.xslt

`metadata/document`

Format	Elementname	Beschreibung	XSLT-Mapping-Datei
PDF	isPDFA	Konform zu PDF/A-ISO-Norm [yes|no]	[FITS-HOME]/xml/exiftool/exiftool_document_to_fits.xslt
	pdfAVersion	PDF-Version [1a | 1b usw.]	[FITS-HOME]/xml/exiftool/exiftool_document_to_fits.xslt
	graphicsCount	Anzahl der Bildobjekte	[FITS-HOME]/xml/jhove/jhove_document_to_fits.xslt
ePub	author	Autor	[FITS-HOME]/xml/fileAnalyzer/snippets/metadata.xsl
	title	Titel	[FITS-HOME]/xml/fileAnalyzer/snippets/metadata.xsl
	language	Sprache	[FITS-HOME]/xml/fileAnalyzer/snippets/metadata.xsl
	isRightsManaged	enthält Einschränkungen bzgl. der Nutzungsrechte	[FITS-HOME]/xml/fileAnalyzer/snippets/metadata.xsl
HTML	usesStyleSheets	Verwendet StyleSheets	[FITS-HOME]/xml/nlnz/fits/nlnz_html_to_fits.xslt
	paragraphCount	Anzahl Absätze	[FITS-HOME]/xml/nlnz/fits/nlnz_html_to_fits.xslt
	characterCount	Zeichenanzahl	[FITS-HOME]/xml/nlnz/fits/nlnz_html_to_fits.xslt
	wordCount	Wortanzahl	[FITS-HOME]/xml/nlnz/fits/nlnz_html_to_fits.xslt
PS	pageCount	Seitenzahl	[FITS-HOME]/xml/fileutility/fileutility_to_fits.xslt
	charSet	Zeichensatz	[FITS-HOME]/xml/fileutility/fileutility_to_fits.xslt
DOC	paragraphCount	Anzahl Absätze	[FITS-HOME]/xml/exiftool/exiftool_document_to_fits.xslt
	lineCount	Zeilenzahl	[FITS-HOME]/xml/exiftool/exiftool_document_to_fits.xslt

metadata/image

Format	Elementname	Beschreibung	XSLT-Mapping-Datei
PNG	bitsPerSample	Bit pro Sample (gemappt aus EXIFs bitDepth)	[FITS-HOME]/xml/exiftool/exiftool_image_to_fits.xslt
	colorSpace	Farbraum (gemappt aus EXIFs colorType)	[FITS-HOME]/xml/exiftool/exiftool_image_to_fits.xslt

metadata/audio

Format	Elementname	Beschreibung	XSLT-Mapping-Datei
WAV	AverageBytes-PerSecond	Durchschn. Zahl der Bytes pro Sekunde (gemappt aus EXIFs bitDepth)	[FITS-HOME]/xml/nlnz/fits/nlnz_wav_to_fits.xslt
MP3	channelMode	Kanalmodus (gemappt aus EXIFs colorType)	[FITS-HOME]/xml/exiftool/exiftool_audio_to_fits.xslt

metadata/video

Format	Elementname	Beschreibung	XSLT-Mapping-Datei
MPEG	duration	Abspieldauer	[FITS-HOME]/xml/exiftool/exiftool_video_to_fits.xslt
	bitRate	Bitrate	[FITS-HOME]/xml/exiftool/exiftool_video_to_fits.xslt
	frameRate	Framerate	[FITS-HOME]/xml/exiftool/exiftool_video_to_fits.xslt
	imageWidth	Bildbreite	[FITS-HOME]/xml/exiftool/exiftool_video_to_fits.xslt
	imageHeight	Bildhöhe	[FITS-HOME]/xml/exiftool/exiftool_video_to_fits.xslt
	blockSizeMin	Min. Blockgröße	[FITS-HOME]/xml/exiftool/exiftool_video_to_fits.xslt
	blockSizeMax	Max. Blockgröße	[FITS-HOME]/xml/exiftool/exiftool_video_to_fits.xslt

Referenzen

[UOF-SIP] Digital Information Archiving System - SIP Interface Specification: http://kopal.langzeitarchivierung.de/downloads/kopal_DIAS_SIP_Interface_Specification.pdf [23.08.2012].

[LMER] *Langzeitarchivierungsmetadaten für elektronische Ressource*: http://www.dnb.de/DE/Standardisierung/LMER/lmer_node.html [23.08.2012].

[UOF] *Universelles Objektformat*: http://kopal.langzeitarchivierung.de/index_objektspezifikation.php.de [23.08.2012].

[FITS] *FITS-Projekthomepage* http://projects.iq.harvard.edu/fits [07.07.2014].

[FITS-Schema] https://github.com/harvard-lts/fits/blob/master/xml/fits_output.xsd [07.07.2014].

[FITS-techElements] https://github.com/carlwilson/fits-old-googlecode/blob/master/xml/fits_output_tech_elements.xsd [07.07.2014].

[FITS-DNB-Schema] http://files.dnb.de/standards/fits/fits_output_dnb.xsd [07.07.2014].