Niedersächsisches Landesarchiv

NLA-Katalog archivfähiger Dateiformate

Information für abgebende Stellen

Stand: 5. April 2023

Dokumentinformationen

Kurztitel	Dateiformatekatalog
Zielgruppe(n)	Mitarbeiter/innen im NLA, Abgebende Stellen
Zugehörige Dokumentationen	
Aktenzeichen	56402/14.3

Änderungsnachweis des Dokuments

Datum	Version	Kurzbeschreibung der Änderung	Autor/in	Status
28.08.2020	1.0	Schlussabstimmung in PG DNLA	Lk	final
04.03.2021	2.0	Überarbeitung der Vorgaben für 2., 3., 4., 5. und 6.	Sd	final
29.03.2023	3.0	Änderungen bei 3. und 4., jpeg2000 wurde entfernt	TG DIMAG	final

Inhaltsverzeichnis

1. Vorbemerkung

Archivfähige Dateiformate sind Formate, die sich aufgrund ihrer Verbreitung, der langfristigen Verfügbarkeit und ihres offenen Standards zur langfristigen Erhaltung von Informationen eignen. Im Folgenden werden die Dateiformate und Spezifikationen aufgelistet, die bei der Aussonderung unterschiedlicher Arten von Daten zu verwenden sind.

Im Rahmen einer Anbietung von aussonderungsreifen digitalen Unterlagen an eine Abteilung des Niedersächsischen Landesarchivs (NLA) und vor einer Aussonderung ist in jedem Fall zunächst Kontakt mit dem NLA aufzunehmen, um das Vorgehen bei einer Abgabe abzustimmen. Für die Übernahme von Formaten, die im Katalog nicht aufgeführt werden, ist eine entsprechende Regelung zwischen abgebender Stelle und NLA zu treffen.

Kontakt: dimag@nla.niedersachsen.de

2. Dateiformate

2.1. Reiner Text

TEXT für reine Textinformationen ohne Formatierung.

Spezifikation	Bevorzugt	Weitere
Dateiendung	.txt	-
Zeichenkodierung	UTF-8 (ohne BOM)	UTF-8 (mit BOM), UTF 16, ISO/IEC 8859-1, ISO/IEC 8859-15
Zeilenumbrüche	Carriage Return Linefeed (CRLF)	-

2.2. Gestalteter Text

PDF/A für gestaltete Textinformationen, ggf. mit graphischen Elementen, die in Textverarbeitungsprogrammen, DMS o.Ä. erstellt wurden.

Spezifikation	Bevorzugt	Weitere
Dateiendung	.pdf	-
Standard	PDF/A-2u (für Dokumente mit kopierbarem Textinhalt und möglichem Bildinhalt), PDF/A-2b (für Dokumente mit nicht kopierbarem Textinhalt und/oder möglichem Bildinhalt)	PDF/A-2a, PDF/A-1, PDF/A-3 (ohne Dateianhänge)
Farbraum	sRGB	Graustufen

2.3. Rasterbilder

Rasterbilder sind durch ihren Aufbau definiert, der in einer rasterförmigen Anordnung von farbigen Bildpunkten besteht. Als Hauptmerkmal von Rasterbildern sind die Bildgröße sowie die Farbtiefe zu nennen.

TIFF für Grafiken und photographische Bilder auf Basis von Pixelraster.

Spezifikation	Bevorzugt	Weitere
Dateiendung	.tif	.tiff
Komprimierung	LZW, CCITT T.6	Baseline
Farbtiefe	48 Bit (16 Bit je Kanal)	24 Bit (8 Bit je Kanal)
Farbraum	sRGB	Graustufen
Sonstiges		
Multipage-TIFF	• Das Zusammenfassen mehrerer bildbeinhaltenden Seiten innerhalb eines TIFF-Dokuments, ist nicht gestattet.	

2.4. Videoaufzeichnungen

Bei Videoaufzeichnungen handelt es sich um codierte Bewegtbilder und gegebenenfalls eine oder mehrere kodierte Audiospuren, die in einem Container zusammengeführt werden.

FFV1 & LPCM für verlustfreie, aber schwach komprimierte Archivierung.

Spezifikation	**Bevorzugt**	**Weitere**
Dateiendung	.mkv	-
Container	Matroska (MKV)	-
Videocodec	FFV1 ver. 3.4[1]	FFV1 ver. 3.*
Audiocodec	Linear Pulse-Code-Modulation (LPCM) (siehe 2.5. Audioaufzeichnungen)	-
Sonstiges		
Codierungs-Parameter	• Die GOP-Size muss 1 betragen.[2] • Die Prüfsumme für einzelne Slices muss aktiviert sein.[3] • Die Werte für Coder und Context sollten 1 betragen.[4] • Ein Multi-pass Encoding wird für die weitere Reduzierung der Dateigröße empfohlen.	

MPEG-4 für begrenzte Qualitätsanforderungen jedoch stärkerer Komprimierung.

Spezifikation	**Bevorzugt**	**Weitere**
Dateiendung	.mp4	-
Container	MPEG-4 part 14	-
Videocodec	H.264 / x264	-
Audiocodec	AC3	AAC
Sonstiges		
Codierungs-Parameter	• Es sollte ein langsames Preset gewählt werden, um die Komprimierung effizienter durchzuführen.[5] • Die Qualitätsstufe sollte den Standardwert nicht überschreiten.[6] (Verlustfrei: 0, Standard: 23, Bereich: 0-51)	

[1] ffmpeg Argument „-level 3.4“ oder „-ver 3.4“
[2] ffmpeg Argument: „-g 1“
[3] ffmpeg Argument: „-slicecrc 1“ oder „-ec 1“
[4] ffmpeg Argument „-coder 1 -context 1“
[5] ffmpeg Argument „-preset slow“ oder sogar „-preset veryslow“
[6] ffmpeg Argument im Bereich von „-crf 17“

2.5. Audioaufzeichnungen

WAVE für verlustfreie, nicht komprimierte Audioaufzeichnungen.

Spezifikation	**Bevorzugt**	**Weitere**
Dateiendung	.wav	-
Audiocodec	Linear Pulse-Code-Modulation (LPCM)	
Abtastrate	48 kHz (analoge Quellen), 44,1 kHz (digitale Quellen)	48 kHz (digitale Quellen)
Quantifizierung	16 Bit	24 Bit (analoge Quellen)
Sonstiges		
Kanalanzahl	• Die Anzahl der Kanäle richtet sich nach der Kanalanzahl des Quelldokuments und ist somit nicht fest vorgegeben.	

2.6. Strukturierte Informationen

Sobald Daten in einer derartigen Form vorliegen, dass sie einer Strukturierung bedürfen, handelt es sich um strukturierte Informationen. Je nach Komplexität der abzubildenden Struktur, eignet sich in der Regel eines der folgenden Dateiformate.

CSV für die tabellarische Abbildung von Daten.

Spezifikation	**Bevorzugt**	**Weitere**
Dateiendung	.csv	-
Zeichenkodierung	UTF-8 (ohne BOM)	UTF-8 (mit BOM), UTF-16, ISO/IEC 8859-1, ISO/IEC 8859-15
Feldtrennzeichen	Semikolon	-
Datensatztrenner	Carriage Return Linefeed (CRLF)	-
Textbegrenzungszeichen	Anführungszeichen (“)	Bei Dateien mit ausschließlich numerischen Feldern ist kein Textbegrenzungszeichen notwendig.
Maskierungszeichen	Anführungszeichen (“) Wird ein Anführungszeichen innerhalb eines Feldinhaltes verwendet, muss es durch ein vorangestelltes Anführungszeichen maskiert werden. D.h. ein als Feldinhalt vorkommendes Anführungszeichen ist zu verdoppeln.	-
Kopfdatensatz	Erste Zeile	Bei nicht vorhandenem Kopfdatensatz ist eine Referenztabelle mitzuliefern.

Zeilenumbruch in Feldinhalten	Zeilenumbrüche innerhalb von Feldinhalten sind bei vorhandenem Textbegrenzungszeichen zulässig.	-
Sonstiges		
Feldbeschreibungen	• Die Feldbeschreibung erfolgt in Absprache mit dem Archiv.	

XML dient der Darstellung komplexer, hierarchisch strukturierter Daten

Spezifikation	**Bevorzugt**	**Weitere**
Dateiendung	.xml	-
Standard	xDOMEA ver. 2.4 (für E-Akten), Standardisierte Schemata (XÖV)	xDOMEA ver. 2.3 (für E-Akten), Abstimmung eigener Schemata
Zeichenkodierung	UTF-8 (ohne BOM)	UTF-8 (mit BOM), UTF-16, ISO/IEC 8859-1, ISO/IEC 8859-15
Sonstiges		
Schema	• Zur Inhalts-Validierung wird ein Schema benötigt, das, sofern nicht standardisiert, als .xsd-Dokument mitgeliefert werden muss.	

JSON ist ein einfaches Datenaustauschformat, das hauptsächlich in der inter-maschinellen Kommunikation eingesetzt wird.

Spezifikation	**Bevorzugt**	**Weitere**
Dateiendung	.json	-
Zeichenkodierung	UTF-8 (ohne BOM)	UTF-8 (mit BOM), UTF-16, ISO/IEC 8859-1, ISO/IEC 8859-15
Sonstiges		
Feldbeschreibung	• Die Feldbeschreibung erfolgt in Absprache mit dem Archiv.	

3. Glossar

AAC Advanced Audio Coding (AAC) ist ein von der Moving Picture Experts Group (MPEG) entwickeltes, verlustbehaftetes Audiodatenkompressionsverfahren, das als Weiterentwicklung von MPEG-2 Multichannel im MPEG-2-Standard spezifiziert wurde.

AC3 Adaptive Transform Coder 3 (AC3) ist ein verlustbehaftetes Audiokompressionsverfahren des Unternehmens Dolby. Die Komprimierung basiert – wie auch bei AAC – auf der Tatsache, dass das menschliche Ohr bestimmte Toninformationen nicht wahrnimmt.

BASELINE Siehe TIFF.

BOM Die Byte Order Mark (BOM) dient als Kennung zur Definition der Byte-Reihenfolge und Kodierungsform in UCS/Unicode-Zeichenketten, insbesondere Textdateien. Im Gegensatz zu UTF-16 oder UTF-32 stellt sich das Problem der Byte-Reihenfolge bei UTF-8 nicht.

CCITT T.6 Die CCITT-Komprimierung der Gruppe 4 ist eine verlustfreie Methode zur Bildkomprimierung, die in Faxgeräten der Gruppe 4 verwendet wird, die im Faxstandard ITU-T T.6 definiert sind. Es wird nur für bitonale (Schwarzweiß-) Bilder verwendet. Die Komprimierung der Gruppe 4 ist in vielen proprietären Bilddateiformaten sowie in standardisierten Formaten wie TIFF, CALS, CIT (Intergraph Raster Type 24) und im PDF-Dokumentformat verfügbar.

CRLF Carriage Return und Line Feed (CRLF) sind Steuerzeichen, die benutzt werden, um einen Zeilenumbruch in einer Textdatei darzustellen. Ein Wagenrücklauf (CR), auf den ein Zeilenvorschub (LF) direkt folgt (CRLF, \r\n, oder 0x0D0A), bewegt den Cursor in die nächste Zeile, danach an den Beginn der Zeile und ist der übliche Zeilenumbruch in Betriebssystemen des Unternehmens Microsoft.

CSV Das Dateiformat CSV wird zur Speicherung oder zum Austausch einfach strukturierter, tabellarischer Daten verwendet. Ein Datensatz ist dabei in der Regel zeilenbasiert abgebildet. Die einzelnen Daten werden durch einen Separator voneinander getrennt, wodurch sich auch der Name Comma-separated values (CSV) ableitet.

DMS Der Begriff Dokumentenmanagementsystem (auch Dokumentenverwaltungssystem) bezeichnet die datenbankgestützte Verwaltung elektronischer Dokumente.

FFMPEG Das FFmpeg-Projekt besteht aus einer Reihe von freien Computerprogrammen und Programmbibliotheken, die digitales Video- und Audiomaterial aufnehmen, konvertieren, senden (streamen) und in verschiedene Containerformate verpacken können. Das Kommandozeilenprogramm ffmpeg dient dazu, von einem Video-, Audio- oder Bildformat zu einem anderen zu konvertieren.

FFV1 Der FFmpeg Videocodec 1 (FFV1), ist ein verlustfreier intra-frame Videocodec. Er ist Teil der freien Codec-Sammlung libavcodec des FFmpeg-Projekts.

GOP-Size Die Group of Pictures (GoP) bzw. Bild(er)gruppe ist eine Gruppe von in Abhängigkeit untereinander kodierten, aufeinanderfolgenden Einzelbildern im Bilderstrom eines Videos.

H.264 Auch unter der Bezeichnung MPEG-4/AVC (Advanced Video Coding) bekannt, bezeichnet H.264 einen H.Standard zur Videokompression. Der Codec ist der zehnte Teil des MPEG-4-Standards (MPEG-4/Part 10, ISO/IEC 14496-10).

ISO/IEC 8859 Die Normenfamilie ISO 8859 definiert in 15 verabschiedeten und einer verworfenen Teilnorm verschiedene 8-Bit-Zeichensätze für die Informationstechnik. Die Teilnorm ISO/IEC 8859-1 steht für „Latin-1, Westeuropäisch“, die Teilnorm ISO/IEC 8859-15 unterscheidet sich nur an 8 Positionen vom Latin-1-Zeichensatz (u.a. wurde das Eurozeichen € hinzugefügt).

ISO/IEC 15444-1 Siehe JPEG2000.

JPEG2000 JPEG 2000 (ISO/IEC-15444) ist ein Grafikformat für Rastergrafiken mit Bildkompression, das auf der diskreten Wavelet-Transformation (DWT) basiert und sowohl verlustfreie als auch verlustbehaftete Kompression ermöglicht. Es untergliedert sich in 13 Unterstandards, wobei ISO 1544-1 den Basisstandard bezeichnet, der diverse Ergänzungen und mögliche Erweiterungen ausschließt.

JSON Die JavaScript Object Notation (JSON) ist ein kompaktes Dateiformat und dient dem Zweck des Datenaustausches von strukturierten Daten zwischen Anwendungen. JSON ist von der Programmiersprache unabhängig und kann beliebig verschachtelt werden.

LPCM Die Puls-Code-Modulation (PCM) ist ein Pulsmodulationsverfahren, das ein zeit- und wertkontinuierliches analoges Signal in ein zeit- und wertdiskretes digitales Signal umsetzt. Bei der linearen Quantisierung sind die Wertebereiche gleichmäßig groß. Diese PCM-Art wird Linear Pulse-Code-Modulation (LPCM) genannt.

LZW Der Lempel-Ziv-Welch-Algorithmus (LZW) ist ein verlustfreies Komprimierungsverfahren. Das für die Komprimierung eingesetzte Wörterbuch, dass häufig vorkommende Zeichenfolgen unter einer Abkürzung ansprechbar macht, wird erst zur Laufzeit generiert, ist daher datenspezifisch und somit eignet sich LZW für jede Form von Daten.

Matroska Hierbei handelt es sich um ein freies Containerformat für Video- (MKV) oder reine Audiodaten (MKA), das eine Vielzahl an Codecs und Untertitelformaten unterstützt. Das Format verwendet ein binäres XML-Format für die Containerbeschreibung, wodurch eine erhöhte Flexibilität und Rückwärtskompatibilität gewährleistet wird.

MKV Siehe Matroska.

MPEG-4 Es handelt sich um einen Standard (ISO/IEC-14496), der unter anderem Verfahren zur Video- und Audiodatenkompression beschreibt. Im archivischen Kontext sind insbesondere die Spezifikationen Teil 3/Audio (AAC), Teil 10/Video (H.264), Teil 14/Dateiformat (MP4) sowie Teil 17/Untertitel relevant, in denen Codecs und Formate spezifiziert sind.

PDF/A Als Teilmenge des Portable Document Formats (PDF) genormt, wurde PDF/A (ISO/IEC-19005) für die Langzeitarchivierung digitaler Dokumente entworfen. Mittlerweile gibt es vier verschiedene Normen, von denen PDF/A-4 die neueste, auf PDF-Version 2 fußende Variante darstellt. Für jede Norm existieren unterschiedliche Konformitätsebenen, in denen Einschränkungen zu Inhalt und Aufbau möglicher Dokumente festgelegt sind.

SRGB Der RGB-Farbraum (für die drei Grundfarben Rot, Grün und Blau) wird für selbstleuchtende Systeme verwendet, die dem Prinzip der additiven Farbmischung unterliegen. Standard-RGB (sRGB) wurde extra für Monitore entwickelt, deren farbgebende Basis drei Leuchtstoffe in den genannten Grundfarben sind.

TIFF Das Tagged Image File Format (TIFF oder auch kurz TIF) ist ein Dateiformat zur verlustfreien Speicherung von Bilddaten, das verschiedene Bittiefen-Stufen unterstützt und Komprimierungsmöglichkeiten anbietet. Aufgrund der vielfältigen Form und möglichen

Komplexität von TIFF-Dateien wurde mit Baseline TIFF in der Format-Spezifikation eine Untermenge geschaffen, die jedes TIFF-fähige Programm verarbeiten können sollte.

UTF-8 Das 8-Bit UCS Transformation Format (UTF-8) ist die am weitesten verbreitete Kodierung für Unicode-Zeichen, bei der jedem Unicode-Zeichen eine speziell kodierte Zeichenkette variabler Länge zugeordnet wird.

VBS Ein Vorgangsbearbeitungssystem ist ein IT-System bzw. eine Softwarelösung, die mit dem Ziel eingesetzt wird, Geschäftsvorfälle vom Beginn der Bearbeitung bis zu ihrem Abschluss IT-gestützt bearbeiten und elektronisch dokumentieren zu können.

WAVE Das WAVE-Dateiformat ist ein Containerformat zur digitalen Speicherung von Audiodaten. Enthalten sind meist sogenannte PCM-Rohdaten, also eine zeit- und wertdiskrete Darstellung des zeitlichen Verlaufs eines Signals. Die Qualität des aufgezeichneten Klangs hängt dann von zwei Werten ab, der Abtastrate (Anzahl der Abtastungen pro Zeiteinheit) und der Auflösung (Bit-Tiefe).

XDOMEA Dokumentenmanagement und elektronische Archivierung im IT-gestützten Geschäftsgang (DOMEA) ist ein XÖV-Standard für die Übermittlung von Akten, Vorgängen und Dokumenten.

XML Die Extensible Markup Language (XML), ist eine Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten im Format einer Textdatei, die sowohl von Menschen als auch von Maschinen lesbar ist. Für den Datenaustausch sollte eine dokumentbegleitende Grammatik (z.B. ein XML-Schema) definiert sein, gegen welches das XML-Dokument validiert werden kann.

XÖV XML in der öffentlichen Verwaltung (XÖV) ist ein Standard für den elektronischen Datenaustausch der öffentlichen Verwaltung auf der Basis von Nachrichten in XML-Syntax und zugehörigen Codelisten und Prozessen.