

Kristina Andraschko

Dateiformate in der Langzeitarchivierung

Handbuch Repositorienmanagement, Hg. v. Blumesberger et al., 2024, S. 197–213
<https://doi.org/10.25364/978390337423212>



Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz, ausgenommen von dieser Lizenz sind Abbildungen, Screenshots und Logos.

Kristina Andraschko, Universität Wien, DLE Raum- und Ressourcenmanagement, kristina.andraschko@univie.ac.at

Zusammenfassung

Die verwendeten Dateiformate spielen eine wesentliche Rolle in der Langzeitspeicherung von digitalen Objekten. Durch die ständige Weiterentwicklung unterschiedlicher Technologien lässt sich immer nur eine aktuelle Empfehlung abgeben, die ihre Gültigkeit schnell verlieren kann. Es lassen sich jedoch Kriterien für Dateiformate definieren, durch die eine lange Speicherung begünstigt wird. Verschiedene Arten von Objekten, wie Videos, Tonaufnahmen, Bilder oder Texte stellen dabei unterschiedliche Anforderungen an die Wahl des geeigneten Dateiformats. Auch wenn ein Format alle Voraussetzungen für eine lange Archivierung erfüllt, ist es notwendig, neue Entwicklungen stets zu verfolgen und vorhandene Dateien möglichst verlustfrei anzupassen. Um digitale Objekte über einen langen Zeitraum aufzubewahren, ist es wichtig zu überlegen, in welcher Form sie gespeichert werden sollen. Nicht alle Dateiformate erweisen sich in dieser Hinsicht als zuverlässig. Der vorliegende Text soll einerseits für die Beurteilung von Formaten wichtiges Wissen vermitteln und andererseits einen Überblick über die aktuell häufig empfohlenen Formate bieten, sodass auch nicht genannte oder neue Formate auf ihre Eignung für die Langzeitarchivierung beurteilt werden können.

Schlagwörter: Dateiformat; Langzeitarchivierung

Abstract

Digital Formats for Long-Term Preservation

Deciding which file formats to use in a repository is a crucial task in the long-term preservation of digital objects. Due to constant changes in technology, one can only give recommendations that can very well be outdated quickly. However, there are criteria for file formats that support long-term-preservation. Various types of objects, such as video, audio, images or texts need different considerations to determine the appropriate file format. Even if a format meets all prerequisites for long-term preservation, it is important to be aware of new developments and to adapt existing file formats if necessary. In order to preserve digital objects over a long period of time, it is important to consider in which form they should be stored. Not all file formats prove to be reliable in this respect. On the one hand, this text is intended to provide important knowledge for the evaluation of formats. On the other hand, it provides an overview of the currently frequently recommended formats so that even formats not mentioned or new formats can be evaluated for their suitability for long-term archiving.

Keywords: File format; digital preservation

1. Formate für die Langzeitarchivierung

In der digitalen Langzeitarchivierung spielen Dateiformate eine wichtige Rolle, da sie den Zeitraum der Nutzbarkeit von Objekten mitbestimmen. Ist das Format nicht mehr aktuell, lässt sich eine Datei im schlimmsten Fall nicht mehr öffnen und ist so für das Archiv unbrauchbar.

Durch neue technische Anforderungen und Möglichkeiten kommen kontinuierlich neue Formate hinzu, während andere ihre Bedeutung verlieren und nicht mehr verwendet werden – bis zu dem Zeitpunkt, an dem sie gar nicht mehr verwendet werden *können*, weil beispielsweise die nötigen Programme fehlen. Solche Entwicklungen können schnell passieren, sodass es eine gewisse Herausforderung darstellt, sich unter den verfügbaren Formaten zurechtzufinden. Erschwerend kommt hinzu, dass im alltäglichen Sprachgebrauch und in unwissenschaftlichen Quellen unterschiedliche Begriffe häufig synonym verwendet werden. So spricht man beispielsweise häufig von PDF, auch wenn das Objekt als PDF/A – eine für die Langzeitarchivierung besser geeignete Unterart von PDF – vorliegt.

Je komplexer der Objekttyp, desto mehr scheiden sich auch die Geister darüber, welches „das beste Format“ für diesen Typus ist.

Aufgrund der Vielschichtigkeit dieser Thematik ist ein Verständnis davon, was Dateiformate sind und worauf man achten sollte, wichtig. Daher wird im ersten Teil des vorliegenden Beitrags die Frage behandelt, was Dateiformate sind und welche Kriterien für die Auswahl von Formaten für die Langzeitarchivierung zu beachten sind. Im zweiten Teil werden aktuelle Formatempfehlungen für unterschiedliche Objekttypen beschrieben und mit Anmerkungen ergänzt. Zur besseren Übersichtlichkeit und für die Benutzung in der Praxis sind die Formatempfehlungen am Ende dieses Abschnitts in einer Tabelle zusammengefasst.

2. Was sind Dateiformate?

Um auch bei neuen oder unbekannten Dateiformaten eine Entscheidung im Hinblick auf die Eignung zur Langzeitspeicherung treffen zu können, ist es wichtig zu verstehen, was Dateiformate überhaupt sind und was diese tun. Die Antwort auf diese Fragestellung ist so komplex, dass sie selbst ganze Bücher füllen kann.¹ Da es

1 Eine verständliche Erklärung der Grundlagen bietet das Einführungskapitel in Gumm, H.-P. (2013), S. 4ff. oder Kersken, S. (2021), S. 55ff. Zu in Bibliotheken bzw. bei der Digitalisierung gebräuchlichen Formaten sei auch der Abschnitt Einstellungen, Formate und spezielle Verfahren empfohlen in: Lang, E. (2019), S. 169ff.

hier nur darum gehen soll, ein generelles Verständnis für die Thematik zu schaffen, wird im Folgenden eine vereinfachte Darstellung vorgenommen.

Damit ein Computer mit unseren digitalen Objekten umgehen kann, werden diese als Signale übertragen und gespeichert. Die Einheiten dieser Signale sind Bytes bzw. Bits, die als sogenannter Bitstream in einer Reihenfolge vorliegen, die von Programmen interpretiert und dadurch in der intendierten Form präsentiert werden können, z. B. als Bild oder Text.

Das Dateiformat gibt vor, wie die Bits im Bitstream einer Datei angeordnet sein müssen, damit ein Betriebssystem sie der richtigen Anwendung zuweisen kann, und wie das Programm weiter damit verfahren soll. Dazu gibt es gewisse Regeln, die die jeweiligen Programme kennen müssen, um mit der Datei zu arbeiten. Kennen sie diese Richtlinien nicht, können die Daten nicht korrekt wiedergegeben werden. Dieses Problem zeigt sich in der Praxis oft in Form von Fehlermeldungen, die vermutlich hinreichend bekannt sind („Diese Datei kann nicht geöffnet werden, dieses Format wird nicht unterstützt.“).

Diese Regeln, die definieren, wie die Daten codiert und angeordnet sein müssen, sind in der Spezifikation eines Dateiformats beschrieben. Es ist übrigens ein wichtiges Kriterium für ein Format in der Langzeitarchivierung, dass diese Spezifikation einsehbar ist.

Wie die Speichermedien selbst, können auch Dateiformate mit der Zeit obsolet werden, wenn es kein Gerät oder Programm mehr gibt, das mit ihnen umgehen und sie richtig interpretieren kann. Ein Objekt kann dadurch für das Archiv unbrauchbar werden. Um der Obsoleszenz vorzubeugen, ist ein Monitoring der verwendeten Formate wichtig, aber bereits bei der Auswahl der zu akzeptierenden Formate sollte dieser Fall mitbedacht werden.

3. Kriterien für die Auswahl von Formaten für die Langzeitarchivierung

Eine Gefahr in der Langzeitarchivierung in Bezug auf die Dateiformate ist also, dass es kein Programm mehr gibt, das in der Lage ist, eine Datei zu öffnen und korrekt darzustellen. Aber auch häufige Migration oder eine sich verändernde Rechtslage können dazu führen, dass Formate – und im schlimmsten Fall die digitalen Objekte selbst – unbrauchbar werden. Daher sollten bei der Wahl geeigneter Dateiformate verschiedene Faktoren in Betracht gezogen werden. Es gibt gewisse Kriterien, die dabei als Orientierung dienen:

- Offenheit
- Verbreitung
- Transparenz/Verlustfreiheit
- Selbstdokumentation/Metadaten
- Freiheit von Kopierschutz und Verschlüsselungen
- Externe Abhängigkeiten/Portabilität
- Patentfreiheit

Im Folgenden werden die Kriterien zusammen mit den Fragen, die man sich bei der Wahl eines geeigneten Datenformats stellen kann, näher erläutert.

Offenheit

Ist eine Spezifikation des Formates zugänglich?

Im Idealfall sollte die Spezifikation eines Dateiformats offen einsehbar und dokumentiert sein. Dadurch wird ermöglicht, ein Programm zum Lesen eines Formats nachzubauen, wenn beispielsweise die ursprüngliche Software nicht mehr verfügbar ist. Zu beachten ist auch, ob das Dateiformat urheberrechtlich geschützt (proprietär) ist. Nicht-proprietäre Dateiformate haben meistens eine offene Dokumentation. Es gibt aber auch proprietäre Formate, deren Spezifikationen durch den Rechteinhaber veröffentlicht wurden.

Verbreitung

Wie weit verbreitet ist das Format? Wird es häufig genutzt oder handelt es sich eher um ein Nischenprodukt? Gibt es ein anderes Format im selben Bereich, das hier deutlich mehr genutzt wird?

Je weiter verbreitet ein Format ist, desto mehr Möglichkeiten gibt es in der Regel, Dateien in diesem Format zu lesen. Eine hohe Verbreitung beugt auch einer Obsoleszenz des Formates vor.

Transparenz/Verlustfreiheit

Welche Codierung wird innerhalb des Formates verwendet? Welcher Komprimierungsalgorithmus wird bei diesem Format eingesetzt?

Für die Langzeitarchivierung geeignete Formate sollten möglichst direkt analysiert werden können, z. B. durch die Lesbarkeit in einem Text-Editor. Der Inhalt sollte nach bestehenden Standards codiert sein, so empfiehlt sich z. B. für Text Unicode oder UTF-8.

Für Bilder, Audios und Videos gibt es verlustfreie und verlustbehaftete Formate, wobei man sich dabei in den meisten Fällen auf den verwendeten Komprimierungsalgorithmus bezieht. In vielen Bereichen ist Datenverlust durchaus hinnehmbar, z. B., wenn eine geringe Dateigröße erzielt werden soll. Das kann der Fall sein, wenn nur wenig Speicherplatz zur Verfügung steht, mit dem sparsam umgegangen werden muss. In der Langzeitarchivierung sind Formate mit möglichst verlustfreier Kompression die bessere Wahl. Eine Speicherung ganz ohne Datenverlust bzw. -komprimierung ist nicht immer möglich, da die Dateigrößen bei multimedialen Objekten in vielen Fällen nicht zu bewältigen wären.

Wenn eine Umwandlung in ein anderes Format notwendig wird, ist ebenso eine verlustfreie Umwandlung zu wählen, sofern eine solche umsetzbar ist.

Selbstdokumentation/Metadaten

Welche Arten von Metadaten werden in diesem Format gespeichert, wie kann darauf zugegriffen werden oder wo werden sie dargestellt?

Verschiedene Dateiformate können in unterschiedlichem Maße zusätzliche Informationen speichern, die die Lesbarkeit und Dokumentation eines Objekts vereinfachen. Technische Metadaten können auch mit unabhängigen Tools generiert und vom digitalen Objekt getrennt gespeichert werden, jedoch ist es für die Zukunft eines Objekts wesentlich günstiger, wenn es diese Daten direkt selbst enthält. Zusätzlich können auch deskriptive Metadaten direkt in einer Datei enthalten sein, was ihre Auffindbarkeit und den Umgang mit ihr langfristig erleichtern kann. Sofern das vorliegende Dateiformat eine Speicherung der Metadaten unterstützt, kann diese automatisch, z. B. direkt durch das Aufnahmegerät, oder manuell erfolgen.

Freiheit von Kopierschutz und Verschlüsselungen

Enthält ein Objekt einen Kopierschutz und/oder ist es verschlüsselt?

Der Gedanke, ein Objekt durch integrierte Mechanismen vor dem Kopieren zu schützen, widerspricht der Langzeitarchivierung grundsätzlich. Um Dateien dauerhaft für die Zukunft zu sichern, wird man sie kopieren müssen. Ein Kopierschutz darf auch nicht ohne Weiteres entfernt oder umgangen werden. Daher sollte sichergestellt werden, dass die zu archivierenden Objekte nach Möglichkeit in einem Format ohne Kopierschutz oder Digital-Rights-Management-Systeme (DRM) vorliegen. Ein Format muss allerdings nicht ausgeschlossen werden, nur weil es die Möglichkeit bietet, einen solchen Kopierschutz zu implementieren.

Externe Abhängigkeiten/Portabilität

Mit welchem Gerät wird die Datei geöffnet? Kann ich sie auf anderen Geräten öffnen? Welche Software wird dafür benötigt und gibt es mehr als ein Programm zum Öffnen dieser Datei?

Für die Langzeitarchivierung geeignete Dateiformate sollten nicht an eine bestimmte Hard- oder Software gebunden sein. Mit Blick auf die Zukunft ist immer denkbar, dass die benötigten Geräte oder Programme nicht mehr zur Verfügung stehen.

Patentfreiheit

Ist ein Format patentiert? Wer ist der Patentinhaber? Welche Firmen stehen dahinter, wie positioniert sich der Patentinhaber zu Forschung und Open-Source-Projekten usw.?

Patente verhindern sowohl die Weiterentwicklung von Software, die mit einem Format in Zusammenhang steht, als auch die Planbarkeit in Bezug auf die Nutzung eines Objekts. So können sich beispielsweise Lizenzbedingungen in eine Richtung ändern, die aus unterschiedlichen Gründen von Archiven nicht mehr toleriert wird. Durch zeitlich begrenzte Lizenzierung oder für die Freischaltung benötigte externe Quellen kann nicht garantiert werden, dass eine Datei oder eine Software nach mehreren Jahren immer noch zur Verfügung stehen.

Durch eine genaue Analyse vor der Auswahl eines Datenformats werden gute Bedingungen geschaffen, um die Daten auch in Zukunft noch verwenden zu können. Dass auch mit Bedacht ausgewählte Formate obsolet werden können und daher die Arbeit hier noch nicht getan ist, wird im Abschnitt „Migration und Emulation“ thematisiert.

4. Dateiformate – aktuelle Empfehlungen

Im Folgenden sind die derzeit empfohlenen Formate für Objekte der Typen Text, Bild, Audio und Video beschrieben, da diese momentan am häufigsten archiviert werden sollen. Die Empfehlungen ergeben sich aus den bereits beschriebenen Kriterien, sowie bestehenden Empfehlungen namhafter Bibliotheken und Institute.² Häufige Dateiendungen der jeweiligen Formate sind dabei angegeben, es sei aber

² Informationsquellen und Anlaufstellen werden in Abschnitt 6 „Nutzen verschiedener Quellen und Vernetzung“ angegeben.

darauf hingewiesen, dass die Dateieindung allein nicht automatisch einen Rückschluss auf das tatsächlich vorliegende Datenformat zulässt.

In diesem Kapitel wird in einigen Bereichen ein breiteres Basiswissen vorausgesetzt. In den Fußnoten finden sich zusätzliche Informationen und Links zur Einführung bzw. Vertiefung in das jeweilige Thema. Für die Formate selbst sei an dieser Stelle auf die Formatbeschreibungen der Library of Congress³ hingewiesen.

Text

Genau genommen sind Textdateien codierte Textzeichen, die nacheinander interpretiert werden. Dem Objekt liegt ein standardisierter Zeichensatz zugrunde, anhand welchem der Inhalt decodiert werden kann.⁴ Zusätzlich kann in einer Textdatei auch die Information zu einer Schriftart gespeichert sein. Hier ist Vorsicht geboten, da diese Schriftarten außerhalb der Datei gespeichert sind. Entscheidet man sich für die Speicherung in Form einer Textdatei, sollten verbreitete und standardisierte Zeichensätze (UTF⁵ oder ASCII) verwendet werden. Geht es nicht nur um den Text, sondern auch um die Darstellungsform, muss die Information zur Schriftart in den Metadaten gespeichert werden.

Generell sollten nur Formate verwendet werden, die die Informationen zu Zeichensatz und Schriftart enthalten.

Meist handelt es sich bei den zu archivierenden Dateien aber nicht um reine Textdokumente. Es kann daher sinnvoll sein, Formate zu wählen, die auch mit einer komplexeren Darstellung von Tabellen und Bildern umgehen können. Auch Spreadsheets und Tabellen können in diese weiter gefasste Definition von Textdateien fallen.

Selbstverständlich ist es auch möglich, Texte als Bilder zu speichern. Umgangssprachlich werden auch solche Dateien als Textdateien bezeichnet.

Empfohlen (Text): unformatierter Text (z. B.: *.txt, *.asc, *.c), **PDF/A** (*.pdf), **PDF/UA** (*.pdf), sowie **XML-basierte Formate** (z. B.: EPUB3 *.epub)

Empfohlen (Tabellen, Spreadsheets): CSV (*.csv)

Bedingt geeignet: PDF (*.pdf) LaTeX, TeX (*.tex)

3 <https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml>

4 Eine gute Einführung in die Zeichencodierung bietet das W3C (World Wide Web Consortium): <https://www.w3.org/International/questions/qa-what-is-encoding.de>

5 Unicode Transformation Format. Der Unicode-Standard findet sich auf der Website des Unicode Consortiums: <https://www.unicode.org/versions/Unicode14.0.0/>

Nicht geeignet: **Word** (*.doc), **Powerpoint** (*.ppt)

Anmerkungen

Unformatierter Text soll nach Möglichkeit als ASCII, UTF-8 oder UTF-16 codiert sein. Bei UTF muss die Byte-Order Mark angegeben sein, die die Reihenfolge beschreibt, in der die Bytes gelesen werden müssen. Bei Bedarf kann auch ISO 8859-1⁶ verwendet werden.

Bei PDF/A handelt es sich um eine Spezialisierung des PDF-Standards, die für die Archivierung empfohlen wird. Verglichen mit dem PDF-Standard wurde hier eine Reduzierung vorgenommen und es wurden Funktionen eliminiert, die der Archivierung nicht förderlich wären. So wurde z. B. die Möglichkeit entfernt, externe Elemente einzubinden.

Eine weitere Spezialisierung des PDF-Standards, PDF/UA, definiert den Umgang mit unterschiedlichen Unterstützungstechnologien und damit den barrierefreien Zugang zu einer Datei.

Bei der Verwendung von auf XML⁷ basierenden Formaten ist darauf zu achten, dass keine externen Links verwendet werden und dass die Buchstabencodierung und die Dokumenttypdefinition (DTD) in der Datei angegeben sind.

LaTeX⁸ bzw. TeX-Dateien sind für die Langzeitarchivierung geeignet, wenn sie die empfohlenen Codierungen verwenden. Zusätzlich sollten aber auch die erzeugte PDF-Datei sowie gegebenenfalls verwendete Pakete z. B. mit Schriftarten gespeichert werden.

In Bezug auf die nicht geeigneten Formate ist festzuhalten, dass zu den beliebten, aber proprietären Formaten aus der Microsoft-Office-Familie zwar teilweise bereits Dateiformatspezifikationen veröffentlicht wurden, diese aber nicht zwingend vollständig sind.

Bild

Es gibt eine große Anzahl von Bild- bzw. Grafikformaten. Viele verfolgen unterschiedliche Zwecke wie schnelle Implementierbarkeit oder möglichst wenig benötigten Speicher. In vielen Fällen enthalten die Formate einen Komprimierungsalgorithmus, mit dem Bildpunkte zusammen gespeichert werden. Je nach Art der

6 <https://www.iso.org/standard/28245.html>

7 XML (Extensible Markup Language) Essentials mit Beispielen: <https://www.w3.org/standards/xml/core>

8 The LaTeX-Project: <https://www.latex-project.org/>

Komprimierung ist die Datei verlustbehaftet oder verlustfrei. Verlustfreiheit bedeutet in den meisten Fällen eine größere Datei und damit mehr benötigten Speicher. Zusätzlich unterscheidet man zwischen Vektor- und Rastergrafiken. Eine Rastergrafik besteht aus Bildpunkten in unterschiedlichen Farben, während eine Vektorgrafik aus grafischen Elementen wie Linien und Pfaden besteht.⁹ Für beide Varianten gibt es unterschiedliche Verwendungen und Dateiformate, weshalb sie im Folgenden separat behandelt werden.

Rastergrafik

Empfohlen: **TIFF** (*.tif), **JPEG2000** (*.jp2), **PNG** (*.png)

Bedingt geeignet: **Digital Negative Format** (*.dng), **GIF** (*.gif), **JPEG/JFIF** (*.jpg), **BMP** (*.bmp)

Nicht geeignet: proprietäre Rohdatenformate verschiedener Kamerahersteller, deren Dateiformatspezifikationen nicht offen sind

Anmerkungen

Viele der angegebenen Formate unterstützen unterschiedliche Stufen der Komprimierungen. Für die Langzeitarchivierung empfiehlt sich, eine möglichst verlustfreie Komprimierung zu wählen.

Bei dem Digital Negative Format handelt es sich um ein Rohdatenformat von Adobe Inc., dessen Dateiformatspezifikation offen ist. Da es mit einem Patent belegt ist, ist es nur bedingt für die Langzeitarchivierung geeignet. Möchte man aber die Rohdaten eines Bildes speichern, empfiehlt sich dieses Format wegen seiner Verbreitung und der Offenheit der Spezifikation.

Vektorgrafik

Empfohlen: **Scalable vector graphics** (*.svg)

Nicht geeignet: **InDesign Grafik** (*.indd), **Photoshop** (*.psd, *.psb), **Encapsulated Postscript** (*.eps, *.epsf, *.ps)

Audio

Eine Audiodatei enthält die digitale Form einer Audioaufzeichnung. Wie bei Bildern wird auch bei Audiodateien häufig ein Algorithmus zur Komprimierung angewendet, um eine geringere Dateigröße zu erzielen. Das kann zu Datenverlusten führen, die in der Langzeitarchivierung vermieden werden sollen.

⁹ Als niederschwelliger Einstieg in die Computergraphik sei hier empfohlen: Eck, D. (2021).

Häufig kommen für Audiodateien sogenannte Containerformate zum Einsatz. Der Container beinhaltet dabei verschiedene Daten, die durch einen Codec entschlüsselt werden. Genau genommen kommen durch Container und Codec zwei Formate zum Einsatz.

Empfohlen: **WAVE** (*.wav), **BMF**, **FLAC** [Codec] (Audiodatei als LPCM)

Bedingt geeignet: **Advanced Audio Coding** (*.mp4), **MP3** (*.mp3)

Nicht geeignet: z. B.: **AIFF** (*.aif), **Windows Media Audio** (*.wma), **Ogg** (*.ogg)

Anmerkungen

Das WAVE Format ist eigentlich ein von Microsoft und IBM entwickeltes proprietäres Format. Da die Dokumentation offen und die Verbreitung sehr hoch sind, hat es sich dennoch als Standard durchgesetzt.

Bei WAVE handelt es sich um ein Containerformat, das für die Langzeitarchivierung im Idealfall eine unkomprimierte Tonspur in LPCM-Codierung¹⁰ enthalten sollte.

Broadcast Wave ist eine Unterkategorie des empfohlenen WAVE Formats, das die Speicherung von Metadaten möglich macht. Es wird überwiegend im Rundfunkbereich eingesetzt.

Video

Eine Videodatei ist komplex, da sie neben auditiven auch visuelle Inhalte enthält. Sie besteht meist aus einer Containerdatei, in der unter anderem die Audio- und Video- Streams enthalten sind. Diese Streams haben ein Format, das durch einen Codec (de-)codiert wird.

Meistens ist mit dem Format jenes der Containerdatei gemeint. Die Bezeichnungen für Container, Videoformat und Codec werden häufig synonym verwendet oder manchmal schlicht verwechselt, was die Informationsbeschaffung erschweren kann.¹¹

10 Pulse-Code Modulation beschreibt ein Verfahren, bei welchem analoge Signale in digitale umgewandelt werden. PCM und die spezielle Form LPCM (Linear Pulse-Code Modulation) werden in den Formatbeschreibungen der Library of Congress geführt: <https://www.loc.gov/preservation/digital/formats/fdd/fdd000016.shtml>, <https://www.loc.gov/preservation/digital/formats/fdd/fdd000011.shtml>

11 Für eine vertiefende Auseinandersetzung mit dem Thema Video eignet sich z. B. Weynand, D. (2016).

Die Überlegung, ob ein Format geeignet ist, muss für Videodateien noch weitergehen als bei den bisher behandelten Objekttypen, da nicht nur der Codec, sondern auch das Containerformat für die Langzeitspeicherung geeignet sein muss. Dass nicht alle Containerformate mit jedem Codec verwendbar sind, macht die Aufgabe, Videodateien längerfristig zu speichern, noch komplexer.

Daher ist es im Falle von Videodateien besonders wichtig, sich bei der Formatwahl an die eingangs definierten Kriterien zu halten und z. B. nur standardisierte und offene Codecs zu verwenden. Unkomprimierte Videodateien erfordern meist zu viel Speicherplatz, weshalb man Codecs mit verlustfreier Kompression wählen sollte.

Betrachtet man die Komplexität von Videodateien, ist es nicht verwunderlich, dass es in diesem Bereich unterschiedliche Meinungen und Richtungen in Bezug auf empfohlene Formate gibt. In den meisten Fällen lässt sich das Format gar nicht so einfach wählen, da manche Aufnahmegeräte bereits ein bestimmtes Format ausgeben oder eine Datei bereits vorliegt. Wenn es die Umstände erlauben, ist es empfehlenswert, auch andere Formate, z. B. die bedingt geeigneten, zu akzeptieren und eine größere Formatanzahl in Kauf zu nehmen. Eine Migration in ein empfohlenes Format ist oft nur eingeschränkt sinnvoll, da sie einen Datenverlust bedeuten kann. Aufgrund dessen und in Hinblick auf die sich schnell entwickelnde Technik in diesem Bereich sollte zu häufiges Migrieren von Videodateien besser vermieden werden.

Empfohlen: **Matroska** (*.mkv) [Container] mit den Codecs: FFV1 (Version 3) und FLAC, **MXF** (*.mxf) [Container]

Bedingt geeignet: **Quick Time** (*.mov) [Container] mit den Codecs: 444 (XQ), 4444 oder 444 HQ, **Motion JPEG 2000** (*.mj2, *.mjp2) [Container] mit Codec JPEG 2000, MPEG-2 [Codec], **MPEG-4** (*.mp4)

Nicht geeignet: **Windows Media Video** (*.wmv) **RealVideo** (*.rm, *.rv), **Flash Video** (*.flv),

Anmerkungen

Die Moving Pictures Experts Group (MPEG) ist als Arbeitsgruppe der ISO und der IEC für die Standardisierung von Video- und Audiokompression verantwortlich. In den Standards der MPEG sind sowohl Videoformate als auch -Codecs beschrieben.¹²

¹² <https://www.mpegstandards.org/standards/>

Von der Library of Congress wird das Containerformat Quick Time zusammen mit den Codecs 4444 (XQ), 4444 oder 444 HQ empfohlen.¹³ Allerdings handelt es sich dabei um ein proprietäres Format der Firma Apple, dessen Dokumentation jedoch offen ist.

Motion JPEG 2000 ist ein Containerformat für den Codec JPEG 2000. Beides ist nicht völlig lizenzfrei. Der Codec JPEG 2000 bietet die Möglichkeit, sowohl verlustbehaftet als auch verlustfrei zu komprimieren.

Da FFD1 eine vergleichbare Kompression bietet und damit häufig die bessere Alternative darstellt, ist JPEG 2000 nur bedingt empfehlenswert.

Neben den genannten Objektarten gibt es eine Vielzahl weiterer, die in der Langzeitarchivierung eine Rolle spielen. So können abhängig vom Forschungsfeld ganz andere Formate relevant sein. Durch die fortschreitende technologische Entwicklung entstehen auch neue Elemente, die archiviert werden müssen. Zunehmend gewinnen dynamische Objekttypen wie Datenbanken, Programme oder Computerspiele für die Langzeitarchivierung an Bedeutung. Ein Format kann dann vielleicht nicht mehr vorgeschrieben werden und der Fokus muss darauf liegen, möglichst viele Informationen und Möglichkeiten zum langen Erhalt zusätzlich zu archivieren. Software muss z. B. mit genauen Angaben zur benötigten Umgebung (Betriebssystem, Hardware) und mitsamt Sourcecode gespeichert werden. Auch das Betriebssystem selbst muss unter Umständen archiviert werden, um sicherzustellen, dass das archivierte Objekt für die Zukunft erhalten werden kann.

5. Migration und Emulation

Eingangs wurde die Schnelllebigkeit digitaler Inhalte und Formate bereits thematisiert. Laufend werden neue Formate und Standards entwickelt und neue Technologien verändern immer öfter den Umgang mit Medien.

Für die Langzeitarchivierung bedeutet das, dass die Arbeit nicht beendet ist, wenn ein geeignetes Format für ein Objekt gewählt wurde. Neue Entwicklungen und ihre Verbreitung unter den Endbenutzer:innen müssen beständig verfolgt werden.

Wenn sich die Dateien in einem bestimmten Format nicht mehr verwenden lassen, weil es z. B. keine Programme mehr dafür gibt, spricht man von einem obsoleten Format. Droht ein Format obsolet zu werden, lässt es sich nicht vermeiden, dass

13 LoC recommended formats statement 2021/2022: <https://www.loc.gov/preservation/resources/rfs/RFS%202021-2022.pdf>

ein Objekt von einem Dateiformat in ein anderes migriert, also umgewandelt werden muss. Dabei muss sichergestellt werden, dass so wenig Datenverlust wie möglich auftritt. Wenn Ausgangs- und Zielformat hinreichend dokumentiert sind, lässt sich anhand der technischen Gegebenheiten entscheiden, ob eine verlustfreie Migration möglich ist. Gerade bei komplexen Objekten, z. B. mit audiovisuellen Inhalten, ist eine Umwandlung ohne Informationsverlust nur schwierig zu realisieren. Zu häufiges Migrieren sollte in diesen Fällen daher vermieden werden.

Als Alternative zur Migration kann immer öfter eine Emulation in Erwägung gezogen werden. Dabei wird ein System nachgebildet, um Inhalte wieder verfügbar zu machen, die auf einem aktuellen System nicht mehr abspielbar sind. Mit einer wachsenden Zahl an dynamischen digitalen Objekten gewinnt die Emulation zunehmend an Bedeutung.

6. Nutzen verschiedener Quellen und Vernetzung

Derzeit herrscht im Wesentlichen eine gewisse Einigkeit bei den empfohlenen Formaten für die Objekttypen Text, Bild und Ton. Bei Videodateien, die mehrere Formattypen umfassen, bilden sich unterschiedliche Präferenzen aus. Ein Abweichen von einem empfohlenen Format kann unter gewissen Voraussetzungen auch sinnvoll sein.

Zum Abschluss sei an dieser Stelle noch einmal darauf hingewiesen, dass sich Empfehlungen für geeignete Formate schnell ändern können und dass das passende Format – besonders bei komplexen multimedialen Objekttypen – von vielen unterschiedlichen Faktoren abhängig ist. Es ist daher notwendig, kontinuierlich Recherche zu betreiben. Viele Bibliotheken und Archive geben regelmäßig Formatempfehlungen oder Statements ab, die eine gute Orientierungshilfe darstellen. Bestehende File Format Registries können zurate gezogen und in die eigenen Archive eingebunden werden. Um vorliegende Formate erkennen und besser beurteilen zu können, lassen sich unterschiedliche Tools einsetzen. Einige hilfreiche Informationsquellen sind im Folgenden gelistet. Diese ersetzen nicht die eigene Recherche, denn auch die genannten Quellen können veralten und bilden daher nur eine Momentaufnahme ab.

Formatempfehlungen und Informationssammlungen

Library of Congress Recommended Formats Statement¹⁴

Die amerikanische Library of Congress gibt jährlich Formatempfehlungen aus, die auf der genannten Website abrufbar sind. Dort finden sich neben Empfehlungen für digitale Objekte auch Angaben zu analogen Werken.

Formatpolicy der Universitätsbibliothek Bern¹⁵

Als Beispiel für die Empfehlungen einer deutschsprachigen Bibliothek sei hier die Formatpolicy der Universitätsbibliothek Bern genannt, die zuletzt 2020 auf der Website der Bibliothek veröffentlicht wurde. Die Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen¹⁶ wird dabei als einschlägige Quelle angegeben.

Sustainability of Digital Formats¹⁷

Die Library of Congress bietet zudem eine große Informationssammlung zu nachhaltigen Formaten in der Langzeitarchivierung, wo auch detaillierte Formatbeschreibungen zu finden sind.

IANUS¹⁸

Das Forschungsdatenzentrum Archäologie und Altertumswissenschaften des Deutschen Archäologischen Instituts enthält Formatbeschreibungen von für die Fachrichtung relevanten Formaten. Neben den auch in dem vorliegenden Kapitel behandelten Objekttypen finden sich dort auch Informationen zu Datenbanksystemen und Geoinformationssystemen. Da die angeführten Formatbeschreibungen besonders detailliert sind, sind die Informationen auch außerhalb der angegebenen Fachrichtung hilfreich.

Tools und File Format Registries

PRONOM¹⁹ – File Format Registry des National Archives of the United Kingdom

14 <https://www.loc.gov/preservation/resources/rfs/>

15 https://www.ub.unibe.ch/unibe/portal/unibiblio/content/e6304/e583799/e791542/e791788/files955270/UBBerndLZAFormatpolicy_ger.pdf

16 <https://kost-ceco.ch/>

17 <https://www.loc.gov/preservation/digital/formats/index.html>

18 <https://ianus-fdz.de/it-empfehlungen/inhalt>

19 <https://www.nationalarchives.gov.uk/PRONOM/Default.aspx>

DROID²⁰ – Formatanalysetool des National Archives of the United Kingdom

JHOVE²¹ – Tool zur Formatanalyse und -validierung

Durch das optimale Ausnutzen solcher bestehenden Informationsquellen sowie durch den Austausch mit anderen Institutionen wird die Beschäftigung mit Dateiformaten für alle erheblich erleichtert und muss keine Aufgabe darstellen, die fast unbewältigbar ist.

7. Empfohlene Dateiformate – Überblick

	Empfohlen	Bedingt geeignet	Nicht empfohlen
Text	unformatierter Text (z. B: *.txt, *.asc, *.c), PDF/A (*.pdf), PDF/UA (*.pdf) sowie XML-basierte Formate (z. B.: EPUB3 *.epub), CSV (*.csv)	PDF (*.pdf), LaTeX, TeX (*.tex)	Word (*.doc), Powerpoint (*.ppt)
Bild	TIFF (*.tif), JPEG2000 (*.jp2), PNG (*.png), Scalable vector graphics (*.svg)	Digital Negative Format (*.dng), GIF (*.gif), JPEG/JFIF (*.jpg), BMP (*.bmp)	proprietäre Rohdatenformate verschiedener Kamerahersteller, deren Dateiformatspezifikationen nicht offen sind, InDesign Grafik (*.indd), Photoshop (*.psd, *.psb), Encapsulated Postscript (*.eps, *.epsf, *.ps)
Audio	WAVE (*.wav), BMF, FLAC [Codec] (Audiodaten als LPCM)	Advanced Audio Coding (*.mp4), MP3 (*.mp3)	AIFF (*.aif), Windows Media Audio (*.wma), Ogg (*.ogg)
Video	Matroska (*.mkv) [Container] mit den Codecs: FFV1 (Version 3) und FLAC, MXF (*.mxf) [Container]	Quick Time (*.mov) [Container] mit den Codecs: 444 (XQ), 4444 oder 444 HQ, Motion JPEG 2000 (*.mj2, *.mjp2) [Container] mit Codec JPEG 2000, MPEG-2 [Codec], MPEG-4 (*.mp4)	Windows Media Video (*.wmv) RealVideo (*.rm, *.rv), Flash Video (*.flv)

20 <https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/>

21 <https://jhone.openpreservation.org/>

Bibliografie

- Drümmer, Olaf; Chang, Bettina (2013): PDF/UA in a Nutshell. Accessible Documents With PDF. Berlin: Association for Digital Document Standards e.V. https://www.pdfa.org/wp-content/until2016_uploads/2013/08/PDFUA-in-a-Nutshell-PDFUA.pdf (abgerufen am 17.05.2022)
- Eck, David (2021): Introduction to Computer Graphics. <https://math.hws.edu/eck/cs424/downloads/graphicsbook-linked.pdf> (abgerufen am 17.05.2022)
- El Idrissi, Bouchra (2019): Long-term Digital Preservation. A Preliminary Study on Software and Format Obsolescence. In: Proceedings of the ArabWIC 6th Annual International Conference Research Track 2019. New York: Association for Computing Machinery, Article 13, pp. 1-6. <https://doi.org/10.1145/3333165.3333178>
- Gumm, Heinz-Peter (2013): Einführung in die Informatik. 10. vollständig überarbeitete Auflage. München: Oldenbourg Verlag.
- Ishida, Richard (2015): Zeichencodierung für Anfänger. Übers. Gunnar Bittersmann (2016). <https://www.w3.org/International/questions/qa-what-is-encoding.de> (abgerufen am 17.05.2022)
- Kersken, Sascha (2021): IT-Handbuch für Fachinformatiker*innen. Der Ausbildungsbegleiter. 10., aktualisierte und überarbeitete Auflage. Bonn: Rheinwerk Computing.
- Lang, Elke; Bohne-Lang, Andreas (2019): Praxishandbuch IT-Grundlagen für Bibliothekare. Berlin/Boston: Walter De Gruyter GmbH.
- Neuroth, Heike; Oßwald, Achim et al. (Hg.) (2010): Nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3. Glückstadt: vwh – Verlag Werner Hülsbusch, Fachverlag für Medientechnik und -wirtschaft.
- Rudnik, Pia (2020): Video: Crashkurs Digitale Langzeitarchivierung – Dateiformate. <https://doi.org/10.5281/zenodo.3985075>
- Schelkens, Peter; Touradj, Ebrahimi et al. (2009): The JPEG 2000 Suite. Chichester, West Sussex, U.K.: Hoboken, N.J.
- SLUB Dresden (2021): Langzeitarchivfähige Dateiformate. Version 2.0, 2021-10-01. https://slubarchiv.slub-dresden.de/fileadmin/groups/slubsite/slubarchiv/SLUBArchiv_langzeitarchivfaehige_Dateiformate_v2.0.pdf (abgerufen am 17.05.2022)
- Watkinson, John (2004): The MPEG Handbook. MPEG-1, MPEG-2, MPEG-3, MPEG-4, 2nd ed. Oxford: Elsevier/Focal Press.
- Weynand, Diana (2016): How Video Works. From Broadcast to the Cloud, 3rd ed. New York, London: Focal Press.

Kristina Andraschko ist IT-Administratorin der DLE Raum- und Ressourcenmanagement an der Universität Wien. Sie erarbeitete 2018 die Formatempfehlungen für das Repositorium zur Sicherung von digitalen Beständen an der Universität Wien (PHAIDRA) im Zuge des Grundlehrgangs Library and Information Studies.