



GEVER
Geschäftsverwaltung für Luzern

Richtlinie Dateiformate für Langzeitarchivierung

Version 1.7 vom 16. Juni 2021
Status: Verabschiedet

1. Einleitung	3
2. Grundlagen.....	4
3. Dateiformate	5
4. Bemerkungen zu einzelnen Anwendungsfällen	8
5. Berechtigungen, Verschlüsselung, digitale Signaturen	9
6. Beeinträchtigungen durch Konvertierung.....	10

Diese Vorgaben entsprechen der Weisung des Staatsarchivs über Mindestanforderungen an Geschäftsverwaltungssysteme vom 1. August 2015, gestützt auf das Luzerner Archivgesetz (SRL Nr. 585).

Inkraftsetzung

Die vorliegende Richtlinie ist für alle GEVER-Installationen der kantonalen Verwaltung verbindlich.

.....
Vincenz Blaser
Staatsschreiber

.....
Irene Saieva
Projektleiterin GEVER

Versionen

Version	Datum	Status	Ersteller
0.1	12.11.2009	Entwurfsvorlage	EgGr
0.2	01.2009	Vernehmlassung	EgGr
1.0	23.2.2010	Verabschiedung	AGR GEVER
1.4	10.7.2014	Überarbeitung / Aktualisierung	LiMa
1.5	1.9.2014	Vernehmlassung	STALU intern: ScJu / EgGr / LiMa
1.51	16.10.2014	Feedback ORT	LiMa / EgGr
1.6	31.10.2014	Verabschiedung durch ORT	Sai
1.7	16.6.2021	Aktualisierung	LiMa

1. Einleitung

Basierend auf seinen gesetzlichen Aufgaben¹ definiert das Staatsarchiv Luzern den für die Luzerner Verwaltung geltenden Standard für archivtaugliche Dateiformate.

Die Richtlinie beschreibt den Umgang des Staatsarchivs mit Dateiformaten im Rahmen der digitalen Langzeitarchivierung.

Der vorliegende Standard soll

- die langfristige Verstehbarkeit und Nutzung der digitalen Unterlagen gewährleisten.
- den Aufwand für die Aussonderung der Unterlagen aus GEVER-Systemen in Grenzen halten.
- zukünftige Aufwände für Erhaltungsmaßnahmen archivierter Unterlagen minimal halten.

Um dies zu erreichen, favorisiert das Staatsarchiv auf der Basis von Best Practice die Einschränkung der Anzahl Formate. Für die Dienststellen hat dies zur Konsequenz, dass archivtaugliche Formate so früh wie möglich im Data Lifecycle erzeugt werden sollen.

Dieser Standard dient in zweierlei Hinsicht als Handreichung:

- Datenproduktion: Im Kontakt mit den Dienststellen legt er dar, welche Formate aus archivischer Sicht im aktiven Lifecycle verwendet werden sollen. Für die Planung und Einführung neuer Systeme ist er bindend. Die dort verwendeten Dateiformate müssen archivtauglich sein bzw. müssen von der anbietepflichtigen Dienststelle oder Organ gemäss dem vorliegenden Standard derart aufbereitet werden können, dass sie im Staatsarchiv ohne zusätzlichen Aufwand archivierbar sind.
- Datenarchivierung: Bei Angeboten digitaler Unterlagen zur Archivierung legt er dar, welche Formate archivtauglich sind und in welche Formate migriert werden soll.

In produktiven Installationen werden weitere Formate verwendet (besonders MS Office), die nicht für die Langzeitarchivierung geeignet sind. Deren dauernde Verstehbarkeit und Funktionalität kann nicht garantiert werden.

¹ [aus Archivgesetz, SRL 585] §5, Absatz 2 [Die öffentlichen Organe] berücksichtigen bei der Beschaffung von Hilfsmitteln für die Unterlagenverwaltung, namentlich von elektronischen Datenverarbeitungssystemen und Geschäftskontrollen, die Bedürfnisse der Archivierung. [aus Archivverordnung, aus SRL 586] §2, Absatz 3 Das Staatsarchiv ist [...] bei der Planung elektronischer Datenverarbeitungssysteme und Geschäftskontrollen frühzeitig beizuziehen.

2. Grundlagen

Der vorliegende Standard wurde im Staatsarchiv erarbeitet. Er basiert weitgehend auf Arbeiten von:

- BAR (schweizerisches Bundesarchiv). Deren Archivtaugliche Dateiformate – Standards für die Archivierung digitaler Unterlagen sind zu finden unter <http://www.bar.admin.ch>, Archivierung > Ablieferung von Unterlagen ans Bundesarchiv (weiterführend betr. Anforderungen etc.)
- KOST (Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen). Deren Katalog archivischer Dateiformate ist zu finden unter https://kost-ceco.ch/cms/kad_main_de.html (weiterführend betr. Hintergründe, Auswahlkriterien etc.)
- Best Practice anderer Archive und internationale / ISO-Standards

3. Dateiformate

Katalog archivischer Dateiformate, die für die Langzeitarchivierung benutzt werden können.

Anwendungsbereich	Archivtaugliche Formate
"Office"-Dokumente (Word, Excel, Power-Point, Outlook)	PDF/A
Text (unstrukturiert)	TXT
Tabellen (strukturierte Daten)	CSV
Relationale Datenbanken	SIARD
Rasterbilder	TIFF (oder PDF/A)
Audio	WAVE
Video	MPEG-4

Allgemeiner Hinweis: Die Liste der akzeptierten Formate ist kurz. Geringe Mengen an Dateien müssen in eines dieser Formate umgewandelt werden. Oft sind jedoch grössere Mengen an Dateien bereits in anderen Formaten vorhanden, z.B. PDF, JPEG, MPEG-2 oder MP3. Das Staatsarchiv bietet in diesen Fällen Beratung zum weiteren Vorgehen an.

Im Folgenden werden die Formate kurz erläutert.

PDF/A

Geeignet für: „Office“ Dokumente

Das Portable Document Format (PDF) eignet sich für die Speicherung von Dokumenten, deren Inhalt auf Druckseiten adäquat wiedergegeben werden kann (zum Beispiel Word-Dokumente, Excel-Dokumente, ...).

Das von der ISO (Internationale Standardisierungsorganisation) standardisierte, offene, nicht-proprietäre Format PDF/A-1 (PDF/Archive ISO-19005-1) basiert auf dem proprietären Format PDF 1.4 ergänzt durch folgende Einschränkungen, welche für die Archivtauglichkeit unverzichtbar sind:

- keine Referenzierung externer Daten (d.h. insbesondere: die im Dokument verwendeten Zeichensätze sind im Dokument eingebettet)
- keine Verschlüsselung und kein Passwortschutz
- kein ausführbarer (Programm-)Code
- keine Multimedia-Inhalte
- einige wenige Metadaten

Von diesem 2006 in Kraft gesetzten Standard gibt es zwei Konformitätsgrade PDF/A-1a und PDF/A1-b. Jedes Dokument, welches dem strengeren Standard PDF/A-1a genügt, genügt auch dem weniger strengen Standard PDF/A-1b.

Die später eingeführte Version PDF/A-2 (ISO-19005-2, entspricht PDF-Version 1.7) eignet sich ebenfalls für die Langzeitarchivierung. Auch hier gibt es mehrere Versionen PDF/A-2a, PDF/A-2b und PDF/A-2u. PDF/A-3 ist nicht geeignet, da es jede Art von eingebetteten Dateien erlaubt.

Das Staatsarchiv akzeptiert den Minimal-Standard PDF/A-1b, aber **empfiehlt als Minimal-Standard PDF/A-2u**.

Generell kann jedes Dokument, das gedruckt werden kann, und dessen Druckversion den Inhalt des Dokuments adäquat wiedergibt, für die Archivierung in PDF/A umgewandelt werden. Einzelne Produkte können auch direkt in PDF/A speichern (z.B. Office 2016).

TXT ("Nur Text" /"plain text"):

Geeignet für: einfache, unstrukturierte Textdaten

Eine unstrukturierte Textdatei ist zur Darstellung reinen Textinhaltes geeignet, der mit minimalen Strukturierungsmöglichkeiten auskommt (Zeilen) und ansonsten keine weitere Struktur- bzw. Darstellungsinformation benötigt, d.h., er enthält keine eingebetteten oder sichtbaren Anweisungen zur Darstellung (Fettdruck, eingerückt, farbig, etc.) oder Informationen zur Struktur (Titel, Abschnitt, Unterabschnitt, Inhaltsverzeichnis, etc.). Es hat sich gezeigt, dass gerade diese in ihrer Art sehr einfachen Dateien sich als am längsten haltbar und verstehbar erweisen.

Beispiele sind einfache E-Mail-Meldungen ("Nur-Text", ohne Anhänge), Logfiles oder Kurzbeschreibungen (readme.txt).

Folgende Zeichensätze sind erlaubt: US-ASCII, ISO 8859-1 und 8859-15 (Latin-1 und Latin-9), Unicode (UTF-8, UTF-16). Dateien in anderen Zeichensätzen sollen in Unicode, vorzugsweise UTF-8, umgewandelt werden.

Anmerkung: Im Gegensatz zu "einfachen" Textdateien enthalten strukturierte Textdateien (XML, HTML, CSV, PS ...) nebst dem eigentlichen Inhalt Markierungen, die den Inhalt strukturieren bzw. Anweisungen zur Darstellung (Formatierung) enthalten. Die Markierungen selbst sind Textteile aus demselben Zeichensatz, z.B. Tags in XML und HTML, Steuercodes in PDF oder das Trennzeichen (häufig ein Komma) bei CSV. Ohne die Zusatzinformation über die Struktur und deren Bedeutung sind derartige Dateien nicht interpretierbar und dürfen nicht als reine Textdatei archiviert werden.

Die im Rahmen der amtlichen Vermessung anfallenden Daten können im (strukturierten Text-) Format INTERLIS1 archiviert werden. Die Archivpakete folgen der «Spezifikation Geo-SIP und Geo-Dossier» des Bundesarchivs und umfassen neben Daten-Dateien (.itf) auch die Modelle (.ili) und weitere Dokumentation.

CSV (Comma Separated Values)

Geeignet für: tabellenartig organisierte Daten

CSV-Daten sind strukturierte Text-Daten. Jede Zeile weist Felder (Kolonnen) auf, die durch ein Komma oder ein anderes ausgewähltes Zeichen getrennt sind. Damit lassen sich Tabellen sehr einfach archivieren.

Applikationen wie Excel oder OpenOffice erlauben das Speichern und Einlesen von CSV-Daten.

Da jedoch in diesem Format keine Formeln und keine Daten-Definitionen festgehalten werden, werden nur gut dokumentierte Daten in CSV-Format ihren Wert behalten.

SIARD (Software Independent Archival of Relational Databases)

Geeignet für: Relationale Datenbanken

Für die Archivierung von relationalen Datenbanken (RDB) eignet sich das SIARD-Format (eCH-0165) in den Versionen 1.0 und 2.1.

Dazu steht die Java-basierte SIARD Suite des Schweizerischen Bundesarchivs unentgeltlich zur Verfügung. Sie extrahiert aus einer RDB die notwendigen Daten (Strukturinformation und Inhalte) und erzeugt daraus eine archivfähige Datei im SIARD Format. Dieselbe Suite erlaubt, die Datenbank und ihre Struktur näher zu beschreiben, d.h. mit Metadaten zu ergänzen (z.B. Bedeutung von Tabellen bzw. Spalten).

Das SIARD-Format birgt jedoch eine hohe Komplexität für die Archivierung wie auch für die spätere Nutzung. Die Anwendung muss deshalb in Zusammenarbeit mit dem Staatsarchiv sorgfältig geplant werden.

TIFF (Tagged Image File Format)

Geeignet für: Bilddaten (Rasterbilddaten)

Das TIFF-Format speichert Bilder als Rastergrafik (auch Pixelgrafik oder Bitmap genannt). Es ist flexibel und erweiterbar. Deshalb sind zur Sicherung der langfristigen Verstehbarkeit gewisse Einschränkungen erforderlich (Baseline TIFF gemäss den Empfehlungen der KOST).

JPEG kann wegen der Komprimierung als Archivformat nicht empfohlen werden. Da das Format aber weit verbreitet und stabil ist, und der Informationsverlust durch die Komprimierung ja bereits bei der Produktion aufgetreten sind, kann es bei der Ablieferung akzeptiert werden. Für bestehende Bildsammlungen (JPEG / JPEG2000) muss der Konvertierungsbedarf abgeklärt werden.

Für Vektorgrafiken soll abgeklärt werden, ob eine Umwandlung in PDF/A oder TIFF sinnvoller ist.

WAVE

Geeignet für: Audiodaten

Aus archivarischer Sicht kann das WAVE-Format zur Verwendung empfohlen werden. Die Audiodateien weisen eine hohe Qualität auf, sind seit langem in Gebrauch, weit verbreitet und weitgehend von spezifischer Hardware und Betriebssystemen unabhängig. Der Nachteil des Formates liegt eindeutig in der Grösse der Dateien.

Für bestehende Audio-Sammlungen (MP3) muss der Konvertierungsbedarf abgeklärt werden.

MPEG-4

Geeignet für: Videodaten

MPEG-4 ist eine Kompressionsmethode und ein dazugehöriges Dateiformat zur Speicherung von digitalen Videodaten (ISO-14496). MPEG-4 komprimiert die Videodaten z.T. erheblich und erreicht höhere Kompressionsraten als beispielsweise MPEG-2.

Empfehlung: MPEG-4 (AVC/AAC) ist der akzeptierte Standard für die Ablieferung von Videodaten. Insbesondere soll dies in neuen Projekten, die Videodaten generieren bzw. verarbeiten, frühzeitig berücksichtigt werden.

Codecs: Video MPEG-4 AVC (auch ITU H.264 bezeichnet), Audio MPEG-4 AAC

Einschränkung: Kein Digital Rights Management verwenden.

MPEG-4 stellt immer einen Kompromiss bezüglich Qualität und Erhaltbarkeit dar und wird empfohlen, wenn Ausgangsmaterial als MPEG-4 vorliegt oder erhebliche Mengen von Videodaten mit geringer Qualitätsanforderung archiviert werden sollen. Im Zweifelsfall vorgängig das Staatsarchiv kontaktieren.

Bemerkung: Transkodierung (Umwandlung eines Formats in ein anderes, z.B. MPEG-2 oder WMF in MPEG-4) sowie Dekomprimierung und anschliessende Neukomprimierung führen aufgrund des verlustbehafteten Komprimierungsalgorithmus von MPEG-4 jedes Mal zu einem Informationsverlust. Vor einer Umwandlung in MPEG-4 von Videodaten, die noch in nicht archivtauglichen Formaten vorliegen (z.B. DVDs), muss die abliefernde Stelle in Zusammenarbeit mit dem Staatsarchiv prüfen, wie genau vorzugehen ist. Bestehende DVDs werden provisorisch in diesem Datei-Format (als Image, nicht auf diesem Datenträger) aufbewahrt.

4. Bemerkungen zu einzelnen Anwendungsfällen

Excel-Tabellen

Wenn bei der Archivierung von Tabellenkalkulationsdaten nur die Werte der Tabelle erhalten werden sollen, können Spreadsheets in CSV konvertiert werden.

Wenn die optische Erscheinung ausreicht, sollen Arbeitsblätter in Tabellenform (Spreadsheets) in PDF/A konvertiert werden. Bereits bei dieser Konvertierung muss jedoch mit Verlusten gerechnet werden, da Spreadsheets in der Regel nicht für den Druck optimiert sind.

Die Bewahrung der Funktionalität von Spreadsheets kann langfristig nicht garantiert werden. Trotz der archivischen Bedenken (proprietäres Format bzw. übermässig komplexe Spezifikation) kann XLSX als vorläufiges Archivformat verwendet werden. Dokumente im Excel-Format (XLS) sollen in das aktuelle Excel-Format XLSX konvertiert werden.

DOCX (und die weiteren OOXML-Formate wie PPTX, XLSX etc.)

OOXML ist ein von Microsoft entwickeltes Office-Format auf der Basis von XML. Ein OOXML-Dokument ist eine ZIP-Datei, die verschiedene XML-Dokumente umfasst, welche den Inhalt und die Struktur des Dokuments beinhalten.

Obwohl es sich in erster Linie um XML handelt, eignet es sich nicht für die Archivierung: Die Darstellungserhaltung bei Konvertierungen ist mangelhaft, da in OOXML die Schriften nicht eingebettet sind. Die Dateien können Makros enthalten, was eine zuverlässige Anzeige der Dokumente ebenfalls stören kann.

E-Mail

Für unformatierte E-Mails ist das Textformat ausreichend (TXT, siehe oben). Beilagen (attachments) sind in die entsprechenden, oben erwähnten Formate zu konvertieren.

Als Alternative bietet sich an, die komplette E-Mail (Text und Attachments, sofern möglich) in eine PDF/A-Datei zu konvertieren.

Scanning

Aktuelle Multifunktionsgeräte bieten bei der Scanning-Funktion häufig eine OCR-Option an. Hier ist das Format PDF/A-2u dem TIFF vorzuziehen, da es neben dem optischen Abbild auch den enthaltenen Text speichern kann.

Für weitergehende Informationen vgl. auch den GEVER-Leitfaden Scanning.

Websites

Sollen nur einzelne Seiten bzw. kleine Websites archiviert werden (Sample), ist eine Konvertierung nach PDF/A in Betracht zu ziehen.

Für die Archivierung von komplexeren HTML-Dateien eignet sich PDF/A hingegen oft nicht, da diese wegen Interaktionen, Scripting und nicht-barrierefreien Designtricks oft auf Druckseiten nicht ohne weiteres adäquat wiedergegeben werden können. Spezialisierte kostenpflichtige Produkte (wie z.B. von www.fossilo.com) können die Inhalte umfangreicherer Websites zumindest teilweise in PDF/A konvertieren.

Eine einfache Lösung für dynamische Websites (website harvesting) existiert hier noch nicht, Teile können – zumindest was die Informationen betrifft – wie Datenbanken behandelt werden.

5. Berechtigungen, Verschlüsselung, digitale Signaturen

Passwortgeschützte Dateien und Dateien mit eingeschränkten Benutzungsrechten (z.B. PDF, aus denen nicht kopiert werden darf etc.) eignen sich nicht für die Langzeitarchivierung.

Verschlüsselte Daten eignen sich nicht für die Langzeitarchivierung, da die Verfügbarkeit der Verschlüsselungsmethoden langfristig nicht garantiert werden kann.

Digitale Signaturen können langfristig nicht garantiert werden. Für die finalen Formate in der Ablage machen sie deshalb keinen Sinn.

6. Beeinträchtigungen durch Konvertierung

Nachfolgende Liste soll mögliche Beeinträchtigungen durch Konvertierung aufzeigen. Die Liste ist nicht abschliessend, sondern von den gewählten Konvertierungstools abhängig.

Sie basiert auf dem Leitfaden des deutschen Bundesarchivs: Aussonderung digitaler Unterlagen und deren Archivierung im Bundesarchiv - Ein Leitfaden. Version 1.2, 2010.

Ausgangsformat	Konvertierung zu	Mögliche Beeinträchtigungen durch Konvertierung
TXT	TXT	<ul style="list-style-type: none"> - Encoding (Zeichensätze) - Seitenaufbau (Anzahl Zeilen/Spalten) und Font (Art/Größe)
PDF	PDF/A	(PDF/A-1 entspricht mehrheitlich PDF V 1.4) <ul style="list-style-type: none"> - Signaturen im pdf werden entfernt - Restriktion (z.B. Passwortschutz): Konvertierung in pdf/a nicht möglich; - Wandlung nicht möglich z.B. bei Präsentationen, Scripting, 3D - Transparenzinformationen gehen verloren
DOC / DOCX	PDF/A	<ul style="list-style-type: none"> - Makros / Makroviren - Referenzen auf externe Ressource - Embedded objects - Nicht vorhandene Module (Silbentrennung, Sprachen) - Nicht vorhandene Fonts - Druckereinstellungen (speichern), Formatänderung Seite - Restriktionen (z.B. Passwortschutz)
XLS / XLSX	PDF/A	<ul style="list-style-type: none"> - Formeln - Druckbereich - mehrere Sheets - siehe DOC
PPT / PPTX	PDF/A	<ul style="list-style-type: none"> - Präsentationsinformationen gehen verloren - Animationen - siehe DOC
MSG	PDF/A	<ul style="list-style-type: none"> - Body: TXT (s.o.), HTML (externe Referenzen) - Zusatzinformationen (Betreff, Follow Up, Priorität) - Attachements (sofern konvertierbar)
HTM / HTML	PDF/A	extern verlinkte Bilder werden nicht berücksichtigt, Multimedia-Inhalte
JPG / JPEG	TIFF	- Auflösung fehlt
GIF	TIFF	- animierte GIF
PNG	TIFF	- Transparenzinformationen
(Video-DVD)	MPEG-4	<ul style="list-style-type: none"> - Navigations-Menü - Komprimierungsverluste
FLV / AVI / MKV etc.	MPEG-4	- Komprimierungsverluste