OSTDATA Forschungsdatendienst für die Ost-, Ostmittel- und Südosteuropaforschung

Empfohlene Dateiformate für Forschungsdaten zur Langzeitarchivierung

Bereits die Auswahl der verwendeten Software beeinflusst das Dateiformat und dessen Nachhaltigkeit. Daher sollte bei der Entscheidung für oder gegen eine bestimmte Software darauf geachtet werden, dass Dateiformate unterstützt werden, die möglichst folgende Eigenschaften aufweisen:

- Weit verbreitet und standardisiert
- Nicht proprietär, also nicht von einer Anwendung oder einem Hersteller abhängig, und mit unterschiedlichen Programmen verwendbar
- Offen dokumentiert mit frei verfügbaren technischen Spezifikationen
- Verlustfreie oder keine Kompression
- Einfach dekodierbar oder sogar unmittelbar lesbar, also nicht durch Kodierung versteckt

Die Dateiformate sind bezüglich der Präferenz hierarchisch von oben nach unten gehend angeordnet.

Dateityp	**Empfohlene Dateiformat**	**Bemerkungen**	**Alternative Dateiformate**
Bilder			
Bild	.tiff/.tif • in der baseline TIFF der Version 6, unkomprimiert	Tagged Image File Format • Gute Akzeptanz, keine Patenteinschränkung oder technische Schutzmechanismen. • Der Begriff Baseline bezieht sich darauf, dass hier eine Untermenge von formatspezifischen Eigenschaften definiert ist, die von allen Computerprogrammen unterstützt werden müssen, damit diese eine TIFF-Datei lesen können. • https://en.wikipedia.org/wiki/TIFF	• JPEG (.jpeg, .jpg) • PNG als Format für die Webdarstellung
Vektorgrafik	• .svg/.svgz	Scalable Vector Graphics • Keine Patenteinschränkung oder technische Schutzmechanismen, breite Akzeptanz, hohe Transparenz. • Ist XML basiert und lässt sich somit auch gut konvertieren. • Bietet sich insbesondere für schematische Darstellungen und Grafiken an. • https://en.wikipedia.org/wiki/Scalable_Vector_Graphics	
Texte			
Text (statisch)	.pdf • in der ISO-Standardform PDF/A	Portable Document Format • Gute Akzeptanz für den Zweck der Langzeitarchivierung, wurde eigens zur verbesserten Archivierbarkeit von PDF-Dokumenten erschaffen. • Wenn möglich sollte auch die Quell-Datei (.docx, odt, etc.) zur generierten PDF veröffentlicht werden. Diese kann die Nachnutzbarkeit im Sinne der direkten	

		Editierbarkeit erleichtern. • https://en.wikipedia.org/wiki/PDF/A	
strukturierter Text	.xml • Auf XML basierende Formate wie .tei, .dta	Extensible Markup Language • Ist vollständig strukturiert. • Die Transparenz von XML ist mittelmäßig, d.h. zwar von Menschen lesbar, aber die Wohlgeformtheit und Validität eines XML Dokuments lässt sich besser maschinell prüfen, z.B. mit dem Tool http://jhove.openpreservation.org/documentation/. • XML lässt sich recht gut in andere visuelle Präsentationsformate überführen (PDF, HTML). • Für eine möglichst einfache Überführung von XML in andere Präsentationsformate sollte auf deren Wohlgeformtheit und Validität geachtet werden (siehe oben). Daher ist hierauf und diesbezüglich auch auf die Konformität mit einem bekannten und verbreiteten Profil / Schema besonders zu achten (TEI, DTA Basisformat). • https://en.wikipedia.org/wiki/XML	• Hypertext Mark-up Language (HTML) (.html)
Text	.rtf	Rich Text Format • Gilt im Gegensatz zu Word-Dokumenten (Doc, Docx), als vollständig strukturiert, ist transparent und von viel Konvertierungssoftware interpretierbar (z.B. als Zwischenformat bei freier ePub-Konvertierungssoftware). • Auch ermöglicht es die bessere Nachnutzbarkeit eines Inhalts verglichen mit Word. Der Verbreitungsgrad des Formats nimmt allerdings in der Tendenz ab. • https://en.wikipedia.org/wiki/Rich_Text_Format	
Text	.txt • Als UTF-8-codiert	Text File • Sehr gute Akzeptanz von allen Betriebssystemen und den meisten Textprogrammen. • Bietet allerdings keine Seitenbeschreibung oder Strukturauszeichnung von Text, ist also nicht mit Office oder DTP Dateien vergleichbar. • https://en.wikipedia.org/wiki/OpenDocument	
Text (Office)	.odt/.fodt	Open Document Format for Office Applications • Das beste Format aus der Familie der Office-Formate. ODT ist ein offener Standard mit hoher Transparenz. Sehr geeignet um die Nachnutzung der eigenen Inhalte zu fördern. • Nur bedingt geeignet zur Langzeitarchivierung, da es häufig von gängiger Identifizierungs-Software / Extraktionssoftware nicht erkannt wird. • https://en.wikipedia.org/wiki/OpenDocument	widely-used proprietary formats, e.g. MS Word (.doc/.docx)

Text (Publikation plus Formeln)	.tex	TeX/LaTex • Sehr gute Dokumentation, offener Standard, Verbreitungsgrad in manchen Disziplinen hoch (nicht unbedingt in den Geisteswissenschaften). • Die Akzeptanz durch Software ist leider nicht groß, da es sich bei TeX Dokumenten selbst um Code handelt, welcher je nach LaTex Prozessor unterschiedlich interpretiert werden kann (Vgl. http://apsr.anu.edu.au/publications/LaTeX-preservation.pdf), es besteht also eine hohe softwareseitige Abhängigkeit von spezialisierter Software. Daher sollte LaTeX und gerade auch TeX immer nur zusätzlich zu einer fertigen Dateiversion (idealerweise PDF/A) archiviert werden. • Lokale Spezifikationen, Makros, beteiligte LaTeX/TeX-Pakete sollten ebenfalls dokumentiert werden, idealerweise mit Verweis auf das Repositorium CTAN, das weltweit relativ häufig gespiegelt wurde und somit eine gewisse Verfügbarkeit sichergestellt ist: http://ctan.org/mirrors • https://en.wikipedia.org/wiki/TeX • https://en.wikipedia.org/wiki/LaTeX	
Tabellen/Datenbanken			
Tabellen	.csv	Comma Separated Values • Comma Separated Values ist ein weit verbreitetes und offenes Austauschformat für tabellarische Inhalte. Es kann von den meisten Tabellenkalkulationsprogrammen sowohl gelesen als auch geschrieben werden. Allerdings eignet sich CSV weder zur Darstellung gestalterischer Eigenschaften (Farben, Fonts, etc) noch zur Darstellung komplexer tabellekalkulatorischen Formeln o.ä. • Da verschiedene Programme unterschiedliche Arten von CSV ausgeben sollte für eine störungsfreie Weiterverwendung stets dokumentiert werden, welche Zeichen als Texttrenner und Feldtrenner verwendet werden. • https://en.wikipedia.org/wiki/Comma-separated_values	
Tabellen	.ods/.fods	• Das beste Format aus der Familie der Office-Formate. ODT ist ein offener Standard mit hoher Transparenz. Sehr geeignet um die Nachnutzung der eigenen Inhalte zu fördern.	
Statistik-Softwareformate / Quantitative Daten in	• SPSS portable format (.por) • STATA (*.dta) • R (*.R) • SAS Transport	Statistik-Softwareformate • Datensätze mit umfangreichen Labels/Metadaten sollten Datensätze so übergeben werden, dass sie mit einem der weit verbreiteten Statistikpakete (SPSS, Stata, R oder SAS) genutzt werden können. Dafür gibt es verschiedene Möglichkeiten: • Daten können in den proprietären Formaten der gängigen Statistikprogramme als	

Tabellen mit umfang-reichen Metadaten	(*.sas) • Weit verbreitete (proprietäre) Formate von Statistikpaketen, wie z.B. SPSS (*.sav), Stata (*.dta)	sogenannte Systemfiles (bspw. SPSS System File) übergeben werden . • Daten können in software-spezifischen portablen Dateiformaten (z.B. SAS transport file) geliefert werden. • Tabulator-, Komma- oder Spalten-getrennte Textdatei (“csv”) mit zusätzlicher Setup-Datei (setup, command oder syntax file für SPSS, Stata, SAS usw.) mit entsprechenden Datendefinitionen (Variablennamen u. –label, fehlenden Werten etc.). Alternativ können die Datendefinitionen auch als DDI-XML file übermittelt werden.	
Relationale Daten-banken	.sql • SQL-Dump/Abfrage	Single Query Language/relationale Datenbank • Für relationale Datenbanken wird eine Archivierung in Form eines Exports der Datenbankinhalte, der Strukturinformationen und weiterer Funktionalitäten in textbasierte Formate empfohlen, welche unabhängig vom verwendeten Datenbankmanagementsystem sind. • Neben den Daten in den Tabellen müssen die Datenbankstrukturdefinitionen wie Attributdatentypen, Relationen zwischen den Tabellen und Formeln zwingend mit archiviert werden. • Unabhängig vom gewählten Archivformat ist es nötig, die grafische Benutzeroberfläche zu dokumentieren, zum Beispiel in Form von Bildschirmfotos oder eines eventuell vorhandenen Benutzerhandbuches. • SQL ist geeignet für die Langzeitarchivierung bei Verwendung eines offiziellen ISO/IEC 9075 Standards, z.B. SQL:2008. Daten, Datenbankstruktur und ein Großteil der Funktionalität bleiben erhalten. Der verwendete Standard muss dokumentiert sein. Speichert: I, S, F. • https://en.wikipedia.org/wiki/SQL	
Relationale Daten-banken	SIARD	Software Independent Archiving of Relational Databases • SIARD ist ein auf XML basierendes Dateiformat, welches vom Schweizerischen Bundesarchiv offiziell zur Langzeitarchivierung entwickelt wurde und zusammen mit einem Softwarepaket – der SIARD-Suite – kostenlos genutzt werden kann. Es erlaubt sowohl den Im- als auch Export in verschiedene Datenbankformate und gehorcht ausschließlich offenen Standards. • https://www.bar.admin.ch/bar/de/home/archivierung/tools---hilfsmittel/siard-suite.html	
Audio			

Audio	.wav	Waveform Audio File Format • WAV wurde von Microsoft und IBM entwickelt und ist offen dokumentiert aber proprietär, aufgrund der sehr weiten Verbreitung aber der Quasistandard. • Die Audiodaten sollten als lineares PCM gespeichert werden. • WAV wird durch Drittanbieter Software unterstützt. • https://en.wikipedia.org/wiki/WAV	MPEG-1 Audio Layer 3 (.mp3), nur wenn die Dateien bereits in diesem Format entstanden sind
Audio	.flac	Free Losless Audio Codec • FLAC ein verlustfrei komprimierender Codec, der offen dokumentiert und frei verfügbar ist. • http://www.data-archive.ac.uk/media/2894/managingsharing.pdf • https://en.wikipedia.org/wiki/FLAC	MPEG-1 Audio Layer 3 (.mp3), nur wenn die Dateien bereits in diesem Format entstanden sind
Video			
Video	.mkv	Matroska • Ein offenes Containerformat, das eine große Bandbreite von Codecs und ergänzenden Inhalten unterstützt. Für die Archivierung können die Codecs FFV1 für Video und FLAC für Audio empfohlen werden. Weitere geeignete Codecs für Matroska sind H.264/MPEG-4 AVC und MPEG-2.	.avi: Audio Video Interleave (AVI) proprietäres, aber einfaches und robustes Format mit großer Verbreitung
Video	.mp4	MPEG-4 • Der unter ISO/IEC 14496 zertifizierte MPEG-4-Standard verwendet den Codec H.264/MPEG-4 AVC. • Wird dieses Format mit dem H.264-Codec verwendet, kann es zur Langzeitarchivierung verwendet werden, wenn dies entweder dem Ursprungsformat der Videodatei entspricht oder verlustfreie Kompression verwendet wird. • https://en.wikipedia.org/wiki/MPEG-4	
Video	.mxf	Material Exchange Format • Von der Library of Congress explizit empfohlener Dateiformatstandard zur Aufnahme von Audio- und Videostreams. Auch wenn das Format auch jedwede Art von weiteren Bitstreams aufnehmen kann, sollte es – laut Experten – als das digitale Equivalent zur Videokassette gesehen werden (Zitiert nach LOC). Es handelt sich um einen offenen, gut dokumentierten und gut archivierbaren Standard. • https://en.wikipedia.org/wiki/Material_Exchange_Format	
Geodaten			

Geodaten	ESRI Shapefile • (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn) GeoTIFF	• Für die Langzeitarchivierung von GIS-Daten ist derzeit noch kein allgemeingültiger Standard etabliert, weswegen auf die Vorgaben zu den in das GIS importierten Daten (z.B. Luftbilder, Geländemodell etc.) verwiesen wird. Es sollen jedoch möglichst keine programminternen Formate, die nicht von anderen GIS Systemen importiert werden können, genutzt werden. Bevorzugt werden sollen: • Vektordaten als Esri Shapefile (shp + shx + dbf) • Rasterdaten als Geo-Tiff	
Geodaten	.kml	Keyhole Mark-up Language	
Geodaten	.csv	Zumindest Längen- und Breitenangaben können auch als .csv-Dateien abgelegt werden	
Software			
Software	als Quellcode (unkompiliert), & Dokumentation	• Das Feld der Archivierung und Nachnutzung von Software ist relativ gering erforscht. Auch die Library of Congress kann bisher auf keine Empfehlungen verweisen. • Grundsätzlich empfiehlt es sich im Falle von selbst geschriebener Software, diese zusammen mit allen technologischen Abhängigkeiten zu dokumentieren und den Code zusammen mit einem möglichst generischen Compilat (d.h. als ausführbares Programm möglichst Betriebssystem-unabhängig) abzulegen. Der Quellcode sollte unbedingt plausible kommentiert werde (evtl. kann auf eine automatische Kommentierung zurückgegriffen werden). Als weitere Form der Dokumentation sind Screencasts denkbar. • Emulation und lauffähige virtuelle Maschinen können auch für eine Archivierung von Software in Betracht gezogen werden (http://www.dnb.de/DE/Wir/Projekte/Laufend/emulationMultimediaObjekte.html, https://www.virtualbox.org/)	
Noten	Music XML	XML-basierter Standard zur Codierung und Bearbeitung von musikalischen Noten. Es handelt sich um einen offenen, weit verbreiteten Standard der von vielen Programmen aus der Branche unterstützt wird.	

Empfehlungen zur Verzeichnisstruktur und Dateibenennung.

- Richten Sie eine **klare Struktur** der Verzeichnisse ein und behalten diese bei
- **Kontrollieren Sie** die **Dateiversionen**
  - Legen Sie obsolete Dateiversionen nach einem Backup separat ab
- **Schaffen Sie Konventionen** zur Dateibenennung
  - Dokumentieren Sie alle Namenskonventionen oder genutzten Abkürzungen (bspw. in ihrem Datenmanagementplan)
  - Beispiele:
    [Sediment]_[Probe]_[Instrument]_[YYYYMMDD].dat
    [Experiment]_[Reagens]_[Instrument]_[YYYYMMDD].csv
    [Experiment]_[Versuchsaufbau]_[Versuchsperson]_[YYYYMMDD].sav
    [Projekt]_[Interview]_[Ort]_[Personen-ID]_[YYYYMMDD].mp4
    etc.
  - Nutzen Sie Datums-/Zeitstempel oder eine separate ID (z. B. v1.0.0) für jede Version
  - Das Datum sollte zu Beginn oder am Ende des Dateinamens stehen, um die Sortierung zu erleichtern
  - Vermeiden Sie Sonderzeichen { } [ ] < > ( ) * % # ‘ ; “ , : ? ! & @ $ ~
  - Nutzen Sie den Unterstrich (_)um Namen zu separieren

Quellen:

- https://wiki.de.dariah.eu/pages/viewpage.action?pageId=38080370
- http://www.data-archive.ac.uk/media/2894/managingsharing.pdf
- https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats
- http://www.digitalpreservation.gov/formats/
- http://guides.library.oregonstate.edu/research-data-services/data-management-types-formats
- http://www.forschungsdaten-bildung.de/formate
- http://www.ianus-fdz.de/it-empfehlungen/dateiformate

Gefördert durch: BSB Bayerische StaatsBibliothek Information in erster Linie DFG Deutsche Forschungsgemeinschaft