

Handreichung: Dateiformate

Ebenso wie die Datenorganisation (siehe Handreichung: Gute Datenorganisation [H7]) erleichtert eine frühzeitige Entscheidung über Dateiformate die effiziente Projektarbeit. Bei dieser Entscheidung sollte Folgendes beachtet werden:

- Es sollten möglichst interoperable Dateiformate verwendet werden. Daten, die mit Spezialsoftware erhoben oder ausgewertet werden, sollten wenn möglich nicht ausschließlich im Dateiformat der Software gespeichert werden, da sie nicht frei zugänglich ist.
- Auf lange Sicht wird es für manche Softwarelösungen keine Updates mehr geben, sodass auch die mit diesen verarbeiteten und abgespeicherten Daten nicht mehr gelesen werden können.
- Zudem gibt es disziplinspezifische Programme, die unerlässlich sind und bleiben werden.

Es gibt Empfehlungen für die Wahl langfristig lesbarer Dateiformate:

- Achten Sie darauf, möglichst **offene Dateiformate** zu wählen, die mit **Open-Source-Software** verarbeitet werden kann.
- Wenn möglich, konvertieren Sie nach Abschluss des Projektes Dateiformate von Spezialsoftware in gängige Standardformate, damit sie möglichst unkompliziert und lange gelesen werden können.
- Bei Text-Dateien empfiehlt sich eine Konvertierung in das PDF-Format, damit die Dateien möglichst lange lesbar sind.

Auf den folgenden Seiten sind empfehlenswerte Dateiformate aufgeführt.

Empfehlungen für unabhängige, langfristig lesbare Formate

Textformate

Bezeichnung	Dateiendung	Archivierungsdauer
PDF/A	.pdf	langfristig
Unformatierter Text	.txt, .c, .cpp, .m, ...	langfristig
XML	.xml, .xsd, ...	langfristig
PDF	.pdf	mittelfristig
HTML	.html, .htm	mittelfristig
Word	.docx	mittelfristig
PowerPoint	.pptx	mittelfristig
LaTeX	.tex	mittelfristig
Programmcode	.c, .rb, .go, ...	mittelfristig
Word	.doc	ungeeignet
PowerPoint	.ppt	ungeeignet

Empfohlene Konvertierungen

- Möglichst Word und PowerPoint Dateien zu PDF/A-1b Dateien konvertieren
- LaTeX and TeX möglichst zu PDF/A konvertieren
- Die Qualität von Konvertierungen muss sorgfältig visuell überprüft werden. Besonders zu beachten sind Formeln, Sonderzeichen, Umlaute, spezielle Fonts, Schreibfehler, Auswählen und Suchen im Text, Tabellen, Farben, transparente Objekte, Kommentare, Vektorgrafiken sowie mehrfache Zeichenebenen

Spreadsheets und Tabellenformate

Bezeichnung	Dateiendung	Archivierungsdauer
Comma oder Tab-begrenzte Textdateien	.csv	langfristig
Excel (Containerformat)	.xlsx	mittelfristig
Excel (Binärformat)	.xls	ungeeignet

Empfohlene Konvertierungen

- Für wichtige Objekte (wie z.B. Figuren) sollte möglichst auch eine Kopie als separate Datei abgespeichert werden
- Tabellen können zu ASCII Text-Dateien konvertiert werden (.csv)

Workspace Speicherung für Matlab, R oder S-Plus

Bezeichnung	Dateiendung	Archivierungsdauer
Textdateien für S-Plus	.sdd	mittelfristig
Matlab	.mat	mittelfristig
R-Dateien	.R, .RData	ungeeignet

Empfohlene Konvertierungen

- Matlab Dateien (.mat) als v7.3 MAT-Dateien abspeichern, weil es dadurch einem HDF5-basierten Standard folgt (HDF5 ist ein offener Standard für Tabellen, Metadaten und komplexe Datenstrukturen).
- Der R-Workspace sollte mit dem Paket rhdf5 in HDF5 gespeichert werden. Die S-Plus Funktion data.dump produziert eine Datei, welche mit der R-Funktion data.restore gelesen werden kann.
- Für komplexe Datenstrukturen ist es meist nicht sinnvoll den Workspace als ASCII zu speichern, weil dies zu schwer lesbaren Dateien führt.
- Wichtige Tabellen im Workspace sollten zusätzlich als CSV-Datei gespeichert werden.

Rastergraphikformate

Bezeichnung	Dateiendung	Archivierungsdauer
Tagged Image File Format (TIFF)	.tif	langfristig
Portable Network Graphics (PNG)	.png	langfristig
JPEG2000	.jpeg, .jpg	langfristig
TIFF komprimiert	.tiff	mittelfristig
Graphics Interchange Format (GIF)	.gif	mittelfristig
JPEG / JFIF	.jpg, .jpeg	mittelfristig
Bitmap (BMP)	.bmp	mittelfristig

Vektorgraphikformate

Scalable Vector Graphics (SVG)	.svg	langfristig
Graphik InDesign	.indd	ungeeignet
Illustrator	.iat	ungeeignet
Encapsulated Postscript	.eps	ungeeignet

Empfohlene Konvertierungen

- Vektorgraphikdateien werden langfristig eher schwieriger zu öffnen sein als Rastergraphikdateien (Bitmaps, GIF, usw.). Auch das Einbetten von Vektorgrafiken in PDF-Dateien ist fehleranfällig.
- Dateien in speziellen Vektorgraphik-Formaten, wie InDesign (.indd) oder illustrator (.ait) sollten möglichst auch als baseline TIFF, PDF / A, SVG oder JPEG Datei abgespeichert werden.
- Die Qualität der Konvertierung muss sorgfältig visuell überprüft werden (Schärfe, Auflösung, Farben, halbdurchsichtige Objekte, Beschriftungen)

Quellennachweise:

KIM Konstanz: <https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/formate-erhalten/> [05.07.2024]

NN, Dateiformate: Information zur Archivierung: <https://forschungsdaten.info/securedl/sdl-eyJ0eXAiOiJKV1QiLCJhbGciOiJIUzI1NiJ9.eyJpYXQiOiJlMjE3MTM3NDk0MSwiZXN1ciI6MCwiZ3JvdXBzIjpbMCwtMV0slmZpbGU0IjmaWxlyWRtaW4va29vcGVyYXRpb25lbi9id2ZkbS9mZG0vNS1maW5kZW4tdW5kLW5hY2hudXR6ZW4vMi1mb3JtYXRRLWVyaGFsdGVuL0ZvcmlhdGluZm9ybWV0aW9uLnBkZiIsInBhZ2UiOiJY3MDk3fQ.NnwjobvNF2M5uubDIQi2KGzbSggX80lgh5uL8gHX2yY/Formatinformation.pdf> [15.06.2022]