Empfehlungen zur Verwendung von Dateiformaten beim Einreichen von Forschungsdaten in HeFDI-Repositorien

Forschende sind im Zuge der Guten Wissenschaftlichen Praxis (GWP) dazu verpflichtet, ihre Forschungsdaten für mindestens 10 Jahre aufzubewahren und damit einhergehend in der Verantwortung, sich Gedanken über die Verwendungsfähigkeit und Lesbarkeit der Daten in diesem Zeitraum und ggf. darüber hinaus zu machen. Dies umfasst eine Reihe von Problematiken, wie z.B. die stetig steigenden Datenmengen, die Entwicklungszyklen von Hard- und Software, die Alterung von Datenträgern, aber auch die Frage nach geeigneten Dateiformaten.

Forschungsdaten liegen, abhängig von Erhebung und Methode, in verschiedenen Dateiformaten vor, die unterschiedlich gut für eine langfristige Nutzung geeignet sind. Für die Sicherung von Daten im wissenschaftlichen Bereich gemäß GWP sollte vor allem auf die Kompatibilität mit verschiedenen Programmen und die verlustfreie Konvertierung in alternative Formate geachtet werden. Weitere gute Kriterien sind Transparenz in Form von Mensch- und Maschinenlesbarkeit sowie die anzunehmende Langzeitstabilität (genormter Standard - bekanntes und verbreitetes Format) eines Dateiformats.

Diese Handreichung empfiehlt Dateiformate zur Ablage in einem Repositorium und nimmt dabei den mittelfristigen Zeitraum von bis zu 10 Jahren in den Blick. Eine Langzeitarchivierung, die den Erhalt der Nutzbarkeit der Daten über 10 Jahre hinaus zum Ziel hat, erfordert eine aufwändige Vorbereitung und aktive Pflege der Daten. Die Auswahl eines geeigneten Dateiformats ist eine gute Vorbereitung, reicht aber allein nicht aus, um eine Eignung für die Langzeitarchivierung zu gewährleisten; ebenfalls deckt die Ablage in einem Forschungsdatenrepositorium nicht die Langzeitarchivierung der Daten ab.

Die folgende Tabelle gibt eine aktuelle Einschätzung der Fachstelle Forschungsdatenmanagement und Datenerhalt an der ETH Zürich (ETH-Bibliothek) zur Eignung von häufig verwendeten Datenformaten. Diese Tabelle basiert auf Erfahrung sowie einer ausführlichen Auswertung von Empfehlungen und Richtlinien internationaler Einrichtungen mit Archivierungsauftrag. In der vierten Spalte stehen Empfehlungen zur Konvertierung in geeignetere Formate. Falls eine Konvertierung zwar möglich ist, aber mit einem geringeren Funktionsumfang oder mit Informationsverlust einhergeht, wird empfohlen, die Daten in beiden Formaten abzulegen. Wenn es nicht möglich ist, ein empfohlenes Dateiformat zu verwenden, sind die Daten voraussichtlich 10 Jahre später nicht mehr nutzbar.

Dateiart	Empfohlen	Bedingt geeignet	Nicht geeignet
Text	• PDF/A (*.pdf, bevorzugte Subtypen 2b und 2u) • Unformatierter Text (*.txt oder Quellcode, usw.) kodiert als ASCII, UTF-8 oder UTF-16 mit Byte Order Mark (BOM) • XML (inklusive XSD/XSL/XHTML, etc.; Schema & Buchstabenkodierung inklusive)	• (*.pdf) mit eingebetteten Fonts • Unformatierter Text (*.txt, .asc, *.c,* .h, *.cpp,* .m, *.py,* .r usw.) (ISO 8859-1 kodiert) • Rich Text Format (*.rtf) • HTML und XML (ohne externe Inhalte) • Word *.docx • PowerPoint *.pptx • LaTeX und TeX (inkl. lizenzfreie Softwarepakete mit Spezialfonts und resultierendes PDF) • OpenDocument Formate (*.odm,* .odt, *.odg,* .odc, *.odf)	• Word *.doc • PowerPoint *.ppt
Spreadsheets und Tabellen	• Komma- oder Tab-begrenzte Text Files (*.csv)	• Excel *.xlsx • OpenDocument Formate (*.odm,* .odt, *.odg,* .odc, *.odf)	• Excel *.xls,* .xlsb **Konvertierung:** in .xlsx
Rohdaten und Workspace		• Unformatierter Text (ASCII-kodiert) • S-Plus (*.sdd) • Matlab (*.mat) ab v7.3 MAT-Datei	• Matlab-Dateien *.mat (binär) **Konvertierung:** HDF5-Format. • R-Dateien *.RData **Konvertierung:**HDF5-

		• Network Common Data Format oder NetCDF (*.nc,* .cdf) • Hierarchical Data Format (HDF5) (*.h5,* .hdf5, *.he5)	Format (mit dem Paket rhadf)
Rastergrafik (Bitmap)	• TIFF (*.tif, unkomprimiert, möglichst TIFF 6.0+) • Portable Network Graphics (*.png, unkomprimiert) • JPEG2000 (*.jp2, verlustfreie Komprimierung) • Digital-Negative-Format (*.dng)	• TIFF (*.tif, komprimiert) • GIF (*.gif) • BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (*.jp2, verlustbehaftete Komprimierung)	
Vektorgrafik	• SVG ohne JavaScript binding (*.svg)		• Grafik InDesign (*.indd), Illustrator (*.ait) • Encapsulated Postscript (*.eps) • Photoshop (*.psd)
CAD	• AutoCAD Drawing (*.dwg) • Drawing Interchange Format, AutoCAD (*.dxf) • Extensible 3D, X3D (*.x3d,* .x3dv, *.x3db)		
Ton, Audio	• WAV (*.wav) (unkomprimiert, pulse-code moduliert)	• Advanced Audio Coding (*.mp4) • MP3 (*.mp3)	
Video[1]	• FFV1 Codec (ab Version 3) in Matroska Container (*.mkv)	• MPEG-2 (*.mpg,*.mpeg)	• Windows Media Video (*.wmv) • QuickTime Movie (*.mov)

		• MP4, heisst auch MPEG-4 Part 14 (*.mp4) • Audio Video Interleave (*.avi) • Motion JPEG 2000 (*.mj2,* .mjp2)

[1]Neben dem Dateiformat (bzw. Containerformat) spielen auch der verwendete Codec und die Kompressionsart eine wichtige Rolle.

Quelle: ETH Zürich, ETH-Bibliothek Archivtaugliche Dateiformate, vereinfacht und kommentiert durch HeFDI, CC-BY 4.0

Weitere Informationen

Für Dateiformate, die entweder nicht in den Empfehlungen auftauchen oder als nicht geeignet bezeichnet werden, ist zunächst zu prüfen, ob als Alternative ein Format aus der Empfehlungsliste genutzt werden kann.

Um den Umgang mit Daten, insbesondere Daten in nicht empfohlenen Dateiformaten, zu erleichtern und einen möglichst langen Erhalt der Nutzbarkeit zu ermöglichen, empfiehlt es sich, eine sogenannte README-Datei zusammen mit den Daten abzulegen. In dieser einfachen Textdatei wird der Kontext der Erstellung der Daten beschrieben, v. a. mit welcher Software (inkl. Version) die Daten erstellt wurden, sowie Informationen zu bestimmten Einstellungen von Messinstrumenten, Codierung, und alle weiteren Hinweise, die helfen können, später Rückschlüsse zu ziehen, wie die Daten genutzt werden können.

Es erhöht die Chancen auf langfristige Nutzbarkeit, wenn eingebettete Objekte (wie z. B. Abbildungen, Tabellen, etc.) zusätzlich als separate Datei abgelegt werden.

Bei der Konvertierung empfiehlt es sich, die Qualität des Ergebnisses sorgfältig visuell zu überprüfen, beispielsweise bei Texten insbesondere die Formeln, Sonderzeichen, Umlaute, speziellen Schriftarten.

Weiterführende Links

Anleitung zum Konvertieren von .docx zu .pdf

Vortrag: Der PDF/A-Standard und seine verschiedenen Versionen