ETH Zürich
ETH-Bibliothek
Fachstelle Digitaler Datenerhalt

data-archive@library.ethz.ch
www.library.ethz.ch/Digitaler-Datenerhalt

Archivtaugliche Dateiformate

Programme und Fileformate verändern sich im Laufe der Zeit, sodass alte Dateien nicht mehr zuverlässig gelesen werden können. Dies erschwert die langfristige Nutzung digitaler Information.

Die Fachstelle Digitaler Datenerhalt der ETH Zürich gibt Ihnen deshalb im vorliegenden Dokument Empfehlungen zur Archivtauglichkeit von Fileformaten. Zudem wird erklärt, wie Sie Ihre Dateien allenfalls in geeignetere Formate konvertieren können und wie Sie mit der Software DROID aus grossen Datensammlungen für die Archivierung ungeeignete Dateien aufspüren können.

Unsere Empfehlungen gelten für die Langzeitarchivierung von Forschungsdaten allgemein und sind nicht zwingend für eine Aufnahme Ihrer Daten in das ETH Data Archiv.

1. Einschätzung verschiedener Dateiformate

1.1 Nutzung beschränkt auf zehn Jahre

Falls Sie Ihre Daten für höchstens zehn Jahre nutzen wollen, empfehlen wir die Formate in der mittleren und der linken Spalte von Tabelle 1. Auch weniger bekannte Formate, die in Ihrem Fachgebiet für diese Art von Daten üblich sind, sind normalerweise geeignet.

Es sollten zudem folgende Punkte beachtet werden:

- Dateien in seltenen Formaten sollten Sie möglichst in übliche Formate konvertieren. Dabei sollten Sie jeweils Original und Kopie archivieren.
- Die Dateien sollten nicht auf andernorts gespeicherte Daten, Zeichensätzen, Formatvorlagen oder Programme verweisen sondern solche zusätzlichen Objekte sollten stattdessen mitarchiviert werden. Wenn solch eine Unabhängigkeit von externen Objekten nicht möglich ist, sollten Sie die bestehenden Abhängigkeiten in einer reinen Textdatei dokumentieren („Readme“). Das Readme legen Sie zusammen mit den Daten ab.
- Dateien sollten nicht passwortgeschützt, verschlüsselt oder komprimiert sein. Falls Sie zwingend Daten verschlüsseln müssen, treffen Sie Vorkehrungen, damit Daten auch nach Ihrem Weggang von einer berechtigten Person geöffnet werden können.
- Verwenden Sie nur Buchstaben, Zahlen, Unterstrich (_) und Bindestrich (-) in den Namen von Ordnern und Dateien, also keine Leerzeichen, Schrägstriche, Umlaute, usw.[1]
- Die Dateinamenerweiterung sollte konsistent mit dem tatsächlichen Dateiformat sein.

1.2 Nutzung für mehr als zehn Jahre

Um Dateien für mehr als zehn Jahre zu nutzen, sollten zunächst auf jeden Fall die oben angegebenen Empfehlungen eingehalten werden. Zusätzlich sollten die Fileformate sehr verbreitet sein, möglichst offenen Standards folgen und nicht proprietär sein. Es gibt jedoch keine Gewähr für die langfristige Nutzung, weil diese von zukünftigen Softwareentwicklungen abhängt.

Für eine Aufbewahrung von mehr als zehn Jahren können wir nur Dateitypen in der linken Kolonne von Tabelle 1 empfehlen, also insbesondere PDF/A, ASCII Text, TIFF, PNG, SVG und JPEG2000. Dabei hängt die zukünftige Lesbarkeit einer Datei auch stark von den benutzten Dateieigenschaften ab: Fortgeschrittene Möglichkeiten eines Formats, wie Video innerhalb einer PDF Datei, werden schwieriger zu lesen sein als die grundlegenden Möglichkeiten des Formats.

ETH Zürich
ETH-Bibliothek
Fachstelle Digitaler Datenerhalt

data-archive@library.ethz.ch
www.library.ethz.ch/Digitaler-Datenerhalt

Die Fachstelle Digitaler Datenerhalt wird die archivierten Fileformate periodisch überprüfen und wird sich bemühen, veraltete Formate möglichst in gebräuchlichere Formate zu konvertieren. Die Originaldatei wird dabei immer mitarchiviert.

Tabelle 1: Einschätzung zur zukünftigen Lesbarkeit einiger gebräuchlicher Dateiformate. Für ausführlichere Informationen verweisen wir auf das Florida Digital Archive[2] sowie auf die Tabelle in Rimkus et al, 2014[3] (Fortsetzung auf nächster Seite).

Dateiart	Geeignet zur Nutzung für mehr als zehn Jahre	Geeignet zur Nutzung beschränkt auf zehn Jahre	Nicht geeignet zur Archivierung
Text	– PDF/A (*.pdf) – Unformatierter Text (*.txt, *.c, *.cpp, *.m, usw.) kodiert als ASCII, UTF-8, oder UTF-16 mit Byte Order Mark – XML (inklusive XSD/XSL/XHTML, etc.; wobei Schema und Buchstabenkodierung explizit im File angegeben werden sollen)	– PDF (*.pdf), wobei die Fonts im PDF File eingebettet sein müssen – Unformatierter Text (*.txt, *.c, *.cpp, *.m, usw.) (ISO 8859-1 kodiert) – Rich Text Format (*.rtf) – HTML (mit DOCTYPE Deklaration) – Word *.docx – PowerPoint *.pptx – LaTeX, TeX (Die ASCII Texte sind langfristig lesbar; allenfalls benutzte lizenzfreie Softwarepakete mit Spezialfonts sollten möglichst mitgeliefert werden) – HTML und XML (Die ASCII Texte sind langfristig lesbar; externe Links möglichst vermeiden) – Programmcode wie *.c, *.cpp, usw. (Die ASCII Texte sind langfristig lesbar; benutzte lizenzfreie Softwarepakete und Libraries sollten möglichst mitgeliefert werden.)	– Word *.doc – PowerPoint *.ppt
Spreadsheets und Tabellen	– Komma- oder Tab-begrenzte Text Files (*.csv)	– Excel *.xlsx (Containerformat)	– Excel *.xls, *.xlsb (binäre Formate)
Workspace Speicherung für Matlab, R oder S-Plus		– Text Dateien für S-Plus (*.sdd). Der ASCII Text ist langfristig nutzbar, die spätere maschinelle Lesbarkeit ist jedoch unsicher. – Matlab *.mat in HDF Format speichern, denn nichttriviale Matlab *.mat ASCII Files können mit load nicht gelesen werden (siehe Tabelle 2).	– Binäre Dateien wie Matlab Dateien *.mat (binär), R Dateien *.RData

ETH Zürich
ETH-Bibliothek
Fachstelle Digitaler Datenerhalt

data-archive@library.ethz.ch
www.library.ethz.ch/Digitaler-Datenerhalt

Tabelle 1 Fortsetzung

Dateiart	Geeignet zur Nutzung für mehr als zehn Jahre	Geeignet zur Nutzung beschränkt auf zehn Jahre	Nicht geeignet zur Archivierung
Rastergrafik (Bitmap)	– TIFF (*.tif) (unkomprimiert, möglichst TIFF 6.0, Part 1: Baseline TIFF) – PNG (unkomprimiert) – JPEG2000 (verlustfreie Komprimierung)	– TIFF (*.tif) (komprimiert) – GIF (*.gif) – BMP (*.bmp) – JPEG/JFIF (*.jpg) – JPEG2000 (verlustbehaftete Komprimierung) (*.jp2)	
Vektorgrafik	– SVG ohne JavaScript binding (*.svg)		– Grafik InDesign (.indd), Illustrator (.ait) – Encapsulated Postscript (EPS)
Ton, Audio	– WAV (*.wav) (unkomprimiert, pulse-code moduliert)	– Advanced Audio Coding (*.mp4) – MP3 (*.mp3)	
Video	– Motion JPEG 2000 (ISO/ IEC15444-4) (*.mj2) – AVI (unkomprimiert, motion JPEG) (*.avi) – QuickTime Movie (unkomprimiert, motion JPEG) (*.mov)	– MPEG-1, MPEG-2 (*.mpg,*.mpeg, in den Container Formaten AVI oder MOV) – MPEG-4 (H.263, H.264) (*.mp4, in den Container Formaten AVI oder MOV)	– Windows Media Video (*.wmv)

2. Empfohlene Konvertierungsmethoden

Empfohlene Konvertierungen sind in Tabelle 2 angegeben. Nützliche Konvertierungen hängen auch davon ab, welche Informationen in den Dateien benötigt werden. So könnten Sie die Tabellen in einem Excel File zu *.csv Text Files konvertieren. Falls jedoch Makros, Formeln oder eingebettete Objekte im Excel File vorhanden sind, verlieren sie diese Informationen.

Sie sollten die Qualität der Konvertierung sorgfältig visuell überprüfen. Originaldatei und konvertierte Datei sollten dann archiviert werden.

Gewisse neuere Filetypen (*.docx, *.xlsx, *.pptx) sind sogenannte Container Dateien. Wenn Sie die Dateinamenerweiterung „.zip“ anhängen, können Sie die einzelnen Komponenten ansehen und geeignete einfachere Dateien auch zusätzlich separat speichern.

ETH Zürich
ETH-Bibliothek
Fachstelle Digitaler Datenerhalt

data-archive@library.ethz.ch
www.library.ethz.ch/Digitaler-Datenerhalt

Tabelle 2: Empfohlene Dateikonvertierungen

Dateiart	Empfohlene Konvertierungen
Text	– Sie sollten Word und PowerPoint Dateien möglichst zu PDF/A-1b Dateien konvertieren. Für Microsoft-Dateien Word oder PowerPoint Dateien sollte dazu gemäss unseren Tests folgende Methode verwendet werden: Die Datei mit Word oder PowerPoint öffnen, dann unter Menu "Datei", „Drucken“ auswählen. Bei Drucker „Adobe PDF“ auswählen. Das Feld „Druckereigenschaften“ anwählen und dort „PDF/A-1b: 2005 (RGB)“ auswählen. Dann Schaltfläche „Drucken“. – LaTeX oder TeX möglichst zu PDF/A konvertieren. – Sie müssen die Qualität von Konvertierungen sorgfältig visuell überprüfen. Achten Sie dabei insbesondere auf Formeln, Sonderzeichen, Umlaute, spezielle Fonts, Textschreibfehler, Auswählen und Suchen im Text, Tabellen, Farben, transparente Objekte, Kommentare, Vektorgraphiken sowie mehrfache Zeichenebenen.
Tabellen	– Excel *.xls Dateien zu *.xlsx konvertieren – Für wichtige eingebettete Objekte (wie z. B. Figuren) sollten sie möglichst auch eine Kopie als separate Datei abspeichern – Tabellen könnten Sie folgendermassen zu ASCII Text Dateien (*.csv) konvertieren: In Excel die einzelnen Blätter als *.csv Datei speichern, in R Tabellen mit write.csv speichern und in S-Plus mit „write.table“ als *.sdd Datei speichern.
Workspace Dump in Matlab, R oder S-Plus	– Matlab *.mat Files als v7.3 MAT Dateien abspeichern (mit save -v7.3 x.mat), weil es dadurch einem HDF5-basierten Standard folgt. (HDF5[4] ist ein offener Standard für Tabellen, Mediadaten und komplexe Datenstrukturen.) – Der R Workspace sollte mit dem Packet rhadf5[5] in HDF5 Format gespeichert werden. Die S-Plus Funktion data.dump produziert ein File welches mit der R-Funktion data.restore[6] gelesen werden kann. – Für komplexe Datenstrukturen ist es meist nicht sinnvoll den Workspace als ASCII zu speichern, weil dies auf schwer lesbare Dateien führt. (Einen solchen ASCII Workspace Dump erhält man in mit save(…, ascii = TRUE), in Matlab mit save file.txt –ascii und in S-Plus mit dump().) – Wichtige Tabellen im Workspace sollten zusätzlich als CSV-Datei gespeichert werden.
Grafik	– Vektorgrafikdateien werden langfristig eher schwieriger zu öffnen sein als Rastergrafikdateien (Bitmaps). Auch das Einbetten von Vektorgrafik in PDF Dateien ist fehleranfällig. Dateien in speziellen Vektorgrafik Formaten, wie InDesign (*.indd) oder Illustrator (*.ait), sollten Sie möglichst auch als baseline TIFF, PDF/A-1b (siehe oben), SVG oder JPG Datei speichern. Sie sollten die Qualität der Konvertierung sorgfältig visuell überprüfen (Schärfe, Auflösung, Farben, halbdurchsichtige Objekte, Beschriftungen).

3. Fileformat Verifikation mit DROID

Die kostenfreie JAVA Applikation DROID[7] erlaubt Ihnen für grosse Dateisammlungen einen Überblick über die benutzten Dateiformate. Zudem können sowohl unbekannte Formate als auch Inkonsistenzen zwischen Inhalt und Dateinamenerweiterung ermittelt werden (Figur 1).

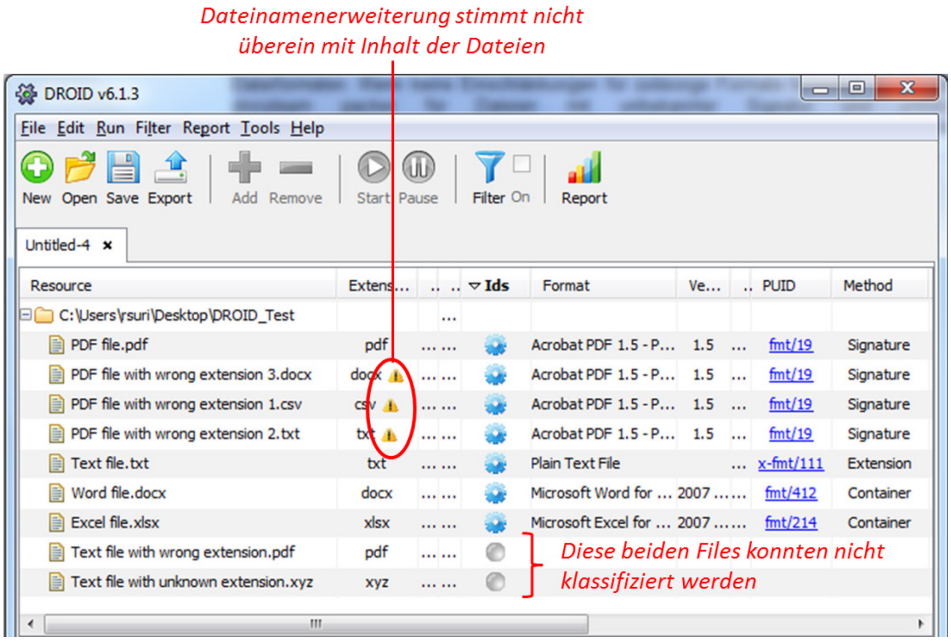
Die meisten Fileformate, ausser den Textdateien, enthalten in den Dateien spezielle Zeichenfolgen um das Dateiformat anzugeben. Diese Zeichenfolgen werden auch Signatur genannt oder „magic numbers“. Falls DROID eine bekannte Signatur innerhalb einer Datei findet, so wird diese Methode benutzt um das Dateiformat zu bestimmen. Unter der Spalte „Method“ (siehe Figur 1) wird dann „Signature“ oder „Container“ angegeben. Falls die Signatur nicht mit der Dateinamenerweiterung übereinstimmt, zeigt DROID eine Warnung (gelbes Dreieck mit Ausrufezeichen).

ETH Zürich
ETH-Bibliothek
Fachstelle Digitaler Datenerhalt

data-archive@library.ethz.ch
www.library.ethz.ch/Digitaler-Datenerhalt

Reine Text Dateien (*.txt) oder auch Tabellen in Text Format (*.csv Dateien) enthalten keine Signatur. DROID klassifiziert solche Dateien nur anhand der Dateinamenerweiterung. Falls keine Signatur gefunden wird und die Dateinamenerweiterung nicht auf ein Textfile hindeutet, so wird die Datei nicht klassifiziert (die untersten beiden Dateien in der Figur).

Die Fachstelle Digitaler Datenerhalt empfiehlt und konfiguriert für gewisse Kunden das Software Tool docuteam packer. Auch docuteam packer findet Dateien mit unklaren oder unbekannten Formaten und erstellt eine Liste analog zu derjenigen von DROID.



Figur 1: Screenshot zur Auswertung einiger Testdateien mit dem Programm DROID. Dateien mit unklaren oder unbekannten Formaten können mit DROID schnell gefunden werden.

Stand und Zugriff auf Quellen: 19. Januar 2015

[1] http://support.apple.com/de-ch/HT5923 (OS X: Bewährte Vorgehensweisen und Methoden für plattformübergreifende Dateinamen, 23.12.2013)

[2] http://fclaweb.fcla.edu/uploads/recFormats_2.pdf ("Recommended Data Formats for Preservation Purposes in the Florida Digital Archive" from Source: FLVC, November 2013)

[3] https://www.ideals.illinois.edu/bitstream/handle/2142/47421/FileFormatStatistics.pdf?sequence=4 (Data from "Digital Preservation File Format Policies of ARL Member Libraries: An Analysis", Kyle Rimkus, Thomas Padilla, Tracy Popp and Greer Martin, D-Lib Magazine, Volume 20, Number 3/4, March/April 2014, doi:10.1045/march2014-rimkus)

[4] http://www.hdfgroup.org/HDF5/doc/UG/UG_frame11Datatypes.html (HDF5 User's Guide, HDF5 Release 1.8.14, November 2014)

[5] http://www.bioconductor.org/packages/release/bioc/vignettes/rhdf5/inst/doc/rhdf5.pdf (rhdf5 - HDF5 interface for R, Bernd Fischer, October 13, 2014)

[6] http://cran.r-project.org/doc/manuals/r-release/R-data.html#EpiInfo-Minitab-SAS-S_002dPLUS-SPSS-Stata-Systat (R Data Import/Export, December 5, 2014)

[7] http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/ (Download DROID: file format identification tool; The National Archives, Version 6.1.5)