Dateiformate für die Langzeitverfügbarkeit

Andreas Frech, Universitätsbibliothek der LMU
Robert Günther, Universitätsbibliothek Bayreuth
Jürgen Rohrwild, Universitätsbibliothek der FAU
Martin Simon, Universitätsbibliothek Regensburg
Martin Spenger, Universitätsbibliothek der LMU
Alexandra Ullrich, Universitätsbibliothek Bayreuth *

Inhaltsverzeichnis

Hintergrund

Die langfristige Verfügbarkeit digitaler Informationen ist nicht nur eine Frage der zur Verfügung stehenden technischen Infrastruktur. Moderne, redundante Archivsysteme können formal den dauerhaften Erhalt von digital erfassten Informationen gewährleisten (Bitstream-Preservation)[1]. Dies erfordert allerdings regelmäßige Investitionen, um alternde Hardware auszutauschen. Damit alleine ist aber noch nicht sichergestellt, dass die Daten langfristig auch tatsächlich nutzbar bleiben. Entscheidend ist auch, ob die digitalen Informationen von zukünftigen Systemen und letztlich Personen noch sinnvoll verstanden werden können.

Dies ist insbesondere dann nicht mehr sichergestellt - und damit die langfristige Verfügbarkeit der Informationen gefährdet - wenn:

*Das Dokument wurde von Martin Spenger und Laura Meier für die Veröffentlichung vorbereitet.
[1]Siehe forschungsdaten.info-Glossar

- die Daten in Dateiformaten abgelegt wurden, die zu einem späteren Zeitpunkt von Computern nicht mehr ohne Informationsverlust interpretiert werden können
- eine spezialisierte, nicht weitverbreitete Software notwendig ist, um die Dateien überhaupt verwenden zu können.

Während der zweite Fall oft bereits zum Zeitpunkt der Archivierung absehbar ist, ist das Risiko eines langsamen „Veraltens“ von Dateiformaten nicht immer offensichtlich. Um dies zu minimieren, sollte idealerweise bereits beim Erstellen der Daten, aber spätestens beim Einbringen von Daten in ein Langzeitspeichersystem darauf geachtet werden, die Informationen in Formaten zu speichern, die allgemein als langfristig sicher angesehen werden. Idealerweise sollte zusätzlich noch berücksichtigt werden, ob das Archivsystem bei drohendem Veralten die Daten automatisch auf modernere Formate migrieren - sprich in aktuelle Datenformate umwandeln - kann.

Je nach erwarteten Nachnutzungsszenarios ist es dabei nicht immer nötig, das aus Sicht der Langzeitstabilität absolut beste Datenformat für den Datentyp zu wählen. Es kann beispielsweise ausreichen, eine pragmatische Lösung zu wählen, die den Erhalt der für die zukünftige Nachnutzungsszenarien entscheidenden Informationen sicherstellt. Dann rechtfertigen die Vorzüge eines formal überlegenden Dateiformats, unter Umständen nicht den mit dem Wechsel des Formates verbundenen Aufwand oder den möglicherweise durch eine Konvertierung entstehenden Informationsverlust.

Die Auswahl eines passenden Dateiformats ist also nicht immer offensichtlich. Daher soll diese Handreichung es Interessierten ermöglichen, sich bereits vor der Einreichung eines Datensatzes bei einem Archiv einen Überblick über geeignete Formate zu verschaffen. Die Empfehlungen wurden ursprünglich im Hinblick auf das in Bayern derzeit genutzte Langzeitarchivierungssystem Rosetta entwickelt. Da die Grundprinzipen der Langzeitverfügbarkeit jedoch unabhängig vom verwendeten Archiv gelten, kann die Handreichung auch in einem breiteren Kontext Anwendung finden.

Anhand realer Fallbeispiele aus verschiedenen Fachdisziplinen mit unterschiedlichen Dateiformaten soll nicht nur eine Empfehlung für das Vorgehen im konkreten Anwendungsfall gegeben werden, sondern auch die sich daraus ergebenden möglichen Konsequenzen aufgezeigt werden. Für einen besseren Überblick werden die konkreten Fallbeispiele durch eine aus bereits existierenden Leitfäden und Handreichungen zusammengetragene Übersicht geeigneter Dateiformate ergänzt.

Der Leitfaden versteht sich ausdrücklich nicht als ein abgeschlossenes Dokument, sondern vielmehr als eine Empfehlungsgrundlage, die kontinuierlich um neue Fallbeispiele sowie weitere Dateiformate ergänzt werden kann.

Aufbau der Fallbeispiele

Der Aufbau der Beispiele folgt grob der Struktur des sogenannten Pattern Konzepts[1]. Dieses wird insbesondere in der Entwicklung verwendet, um kurz gesagt ein auftretendes Problem zu beschreiben und anschließend eine geeignete Lösungsmöglichkeit aufzuzeigen[2]. Dabei wird stets eine fest formalisierte Struktur eingehalten. Diese erlaubt es, die Sammlung an Beispielen einfach und stetig zu erweitern sowie bei Recherchen schneller ein passendes Beispiel für die eigene Fragestellung zu finden. Je nach Anwendungsfall kann es daher für ein und dasselbe Dateiformat

unterschiedliche geeignete Möglichkeiten für die Langzeitverfügbarkeit geben.

Name

Kurze, eindeutige Bezeichnung des Beispielfalls.

Use Case

Beschreibung des konkreten Beispielfalls.

Kontext

Auflistung von für den Beispielfall relevanten Rahmenbedingungen. Kann die Angabe des Datentyps, der Fachrichtung, des geplanten Nachnutzungsszenarios, des Datenvolumen oder weiterer Faktoren sein. Der Kontext soll es erleichtern ähnliche Beispiele zur eigenen Situation zu finden.

Grund

Begründung des empfohlenen Vorgehens.

Konsequenzen und Kosten

Erläuterung des entstehenden Aufwands für die Konvertierung in ein anderes Format und Aufzeigen möglicher Konsequenzen, wenn eine andere Vorgehensweise gewählt wird.

Weitere Hinweise

Verweise auf weiterführende Empfehlungen und Literatur.

Beispielsammlung

Die Beispiele werden im offen zugänglichen Repositorium GitHub gesammelt, kuratiert und bereitgestellt. Dadurch wird zum einen die freie Verfügbarkeit sichergestellt und zum anderen können somit neue Beispielfälle sowie Verbesserungvorschläge und Korrekturen einfach eingebracht werden.

Hier ist der Link zum Repositorium auf GitHub.

Ziel ist es, das Portfolio langfristig weiter auszubauen, mit weiteren beispielhaften Anwendungsfällen anzureichern und diese nachhaltig zur Verfügung zu stellen. Um die Sichtbarkeit zu erhöhen und eine bessere Zugänglichkeit zu erreichen, werden die Beispiele zusätzlich auf Zenodo veröffentlicht sowie in die Webseite der LZV-Initiative Bayern eingebunden.

Allgemeine Empfehlungen

Die Frage der Langzeitverfügbarkeit rückt oft erst beim Abschluss eines Forschungsprojekts in den Fokus. Effizienter ist es jedoch, das Thema bereits frühzeitig bei einzelnen Schritten im Forschungsprozess zu berücksichtigen. Um diese Interventionspunkte und mögliche Fragestellungen nicht zu übersehen, wurde in „3D Data Creation to Curation: Community Standards for 3D Data Preservation“[3] eine hilfreiche und handliche Grafik entworfen, die in Abbildung 1 leicht angepasst und übersetzt wiedergegeben wird.

Abbildung 1: Interventionspunkte und mögliche Fragestellungen zur Langzeitverfügbarkeit

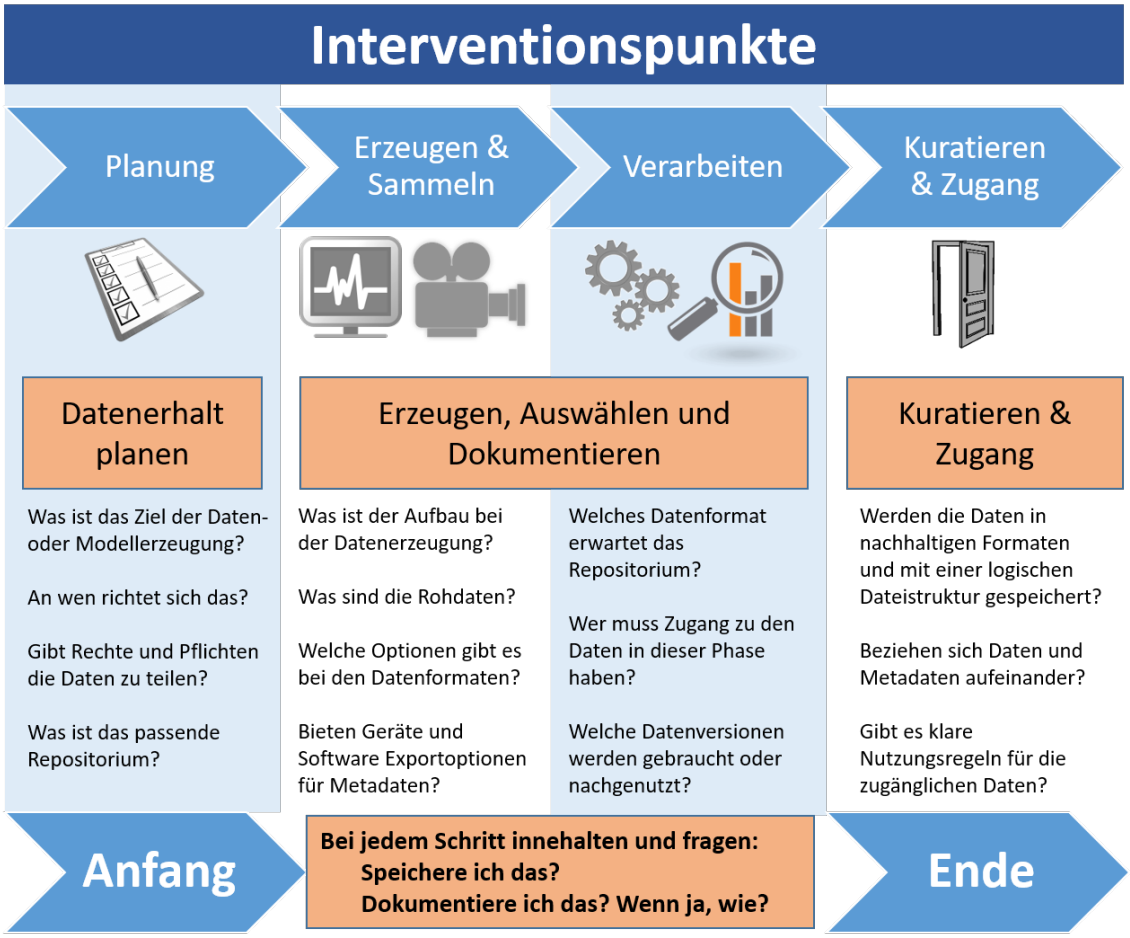

Bild auf Basis von Abbildung 2.1 in Golubiewski-Davis et al. „Best Practices for 3D Data Preservation.” in “3D Data Creation to Curation: Community Standards for 3D Data Preservation”, 2022. CC BY-NC.

Für einen ersten Überblick zu geeigneten Dateiformaten für die Langzeitverfügbarkeit, kann die „Interaktive Tafel gängiger Dateiformate“ der Landesinitiative Langzeitverfügbarkeit lzv.nrw genutzt werden. Anhand der Dateiendung kann schnell überprüft werden, ob das vorliegende Dateiformat für die Langzeitverfügbarkeit geeignet ist.

Weitergehende, detailliertere Erläuterungen zu den einzelnen Dateiformaten finden sich in zahlreichen Quellen[2]. Aus diesen wurde eine kurze Übersicht über empfohlene, bedingt geeignete und nicht geeignete Dateiformate zusammengestellt.

Generell sollte bei der Auswahl von Dateiformaten darauf geachtet werden, dass diese möglichst und je nach Verwendungszweck offen, transparent, weit verbreitet sowie gut dokumentiert sind.

Tabelle 1: Dateiformate und ihre Eignung zur Langzeitverfügbarkeit

Dateityp	**Empfohlene Dateiformate**	**Bedingt geeignete Dateiformate**	**Nicht geeignete Dateiformate**
Text	Unformatierter Text (.txt): kodiert als ASCII, UTF-8 oder UTF-16 mit Byte Order Mark PDF/A (.pdf): explizit für die Langzeitarchivierung entwickeltes Format, die Verwendung der Version PDF/A-2 wird empfohlen XML (inklusive XSD/XSL/XHTML) (.xml): Angabe von Schema und Buchstabenkodierung	PDF (.pdf): spezielle Funktionalitäten und eingebettete Objekte problematisch HTML (.html): Sicherung weiterer Ressourcen zur korrekten Darstellung notwendig Open Document Format (.odt; .odp): offener Standard, nicht eingebettete Elemente und Makros problematisch Office Open XML (.docx; .pptx): proprietäres Format, nicht eingebettete Elemente problematisch LaTeX, TeX (.tex): verwendete Softwarepakete zusätzlich archivieren, nach Möglichkeit nur zusätzlich zu PDF/A-2	Microsoft Word (.doc): obsoletes, proprietäres Format Microsoft PowerPoint (.ppt): obsoletes, proprietäres Format

[2]Siehe forschungsdaten.info, KOST, IANUS

Dateityp	Empfohlene Dateiformate	Bedingt geeignete Dateiformate	Nicht geeignete Dateiformate
Rastergrafik	Tagged Image File Format (TIFF)(.tiff): keine Patenteinschränkung oder technische Schutzmechanismen Portable Network Graphics (PNG)(.png): verlustfrei komprimierendes Format u.a. für die Webdarstellung JPEG2000 (.jp2, .jpg2): verlustfrei komprimierendes Format, v.a für Fotografien	JPEG (.jpeg, .jpg): verlustbehaftete Komprimierung Graphics Interchange Format (GIF)(.gif): Vorgänger des PNG-Formats	
Vektorgrafik	Scalable Vector Graphics (.svg; .svgz): XML-basiertes offenes Format, ermöglicht einfache Zugänglichkeit, nach Möglichkeit ohne script bindings archivieren	Drawing Interchange File Format (.dxf): offen dokumentiertes, aber proprietäres Austauschformat für CAD-Daten	Adobe Illustrator (.ai): proprietäres Format CorelDraw (.cdr): proprietäres Format
Tabellen	Comma-separated values (CSV)(.csv): weit verbreitetes, offenes Austauschformat mit reinen Textdateien	Open Document Spreadsheet (ODS)(.ods): offener Standard, falls Erhaltung der Funktionalität notwendig, nicht eingebettete Inhalte problematisch Office Open XML (.xlsx): proprietäres Format, falls Erhaltung der Funktionalität notwendig	Microsoft Excel (.xls): obsoletes, proprietäres Format

Dateityp	Empfohlene Dateiformate	Bedingt geeignete Dateiformate	Nicht geeignete Dateiformate
Datenbanken	Software Independent Archival of Relational Databases (SIARD)(.siard): auf XML-basierendes offenes Format für die Langzeitarchivierung, Archivierung der Datenbankstruktur und -inhalte Strucured Query Language (SQL)(.sql): SQL-Dump, geeignet insofern ein offizieller, dokumentierter ISO-Standard verwendet wird	Comma-separated values (CSV)(.csv): erlaubt keine Darstellung von Beziehungen, Metadaten und Strukturinformationen	Microsoft Access (.mdb): obsoletes, proprietäres Format
Audio	Waveform Audio File Format (WAV)(.wav): weit verbreitetes, offenes Containerformat für unkomprimierte Audiodaten Free Lossless Audio Codec (FLAC)(.flac): frei verfügbarer, verlustfrei komprimierender Codec	MPEG-1 Audio Layer 3 (MP3)(.mp3): weit verbreitetes, verlustbehaftet komprimierendes Austauschformat	

Dateityp	Empfohlene Dateiformate	Bedingt geeignete Dateiformate	Nicht geeignete Dateiformate
Video	Matroska (.mkv): offenes Containerformat, das zahlreiche Codecs unterstützt, in Verbindung z.B. mit dem Codec FFV1 für die LZA geeignet	MPEG-4 Part 14(MP4)(.mp4): ISO-zertifiziertes Containerformat, kann in Verbindung mit dem H.264-Codec sowie einer verlustfreien Kompression verwendet werden Motion JPEG 2000 (MJ2): ISO-zertifiziertes Containerformat, sollte nur in Verbindung mit verlustfrei komprimierenden Codecs verwendet werden	Audio Video Interleave (AVI)(.avi): proprietäres Format Quick Time File Format (MOV)(.mov): proprietäres Containerformat

Literatur

[1] C. Alexander, S. Ishikawa, M. Silverstein, and et al. *A Pattern Language*. Oxford University Press, New York, 1977.

[2] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design patterns: Entwurfsmuster als Elemente wiederverwendbarer objektorientierter Software*. mitp, Heidelberg, 2015.

[3] K. Golubiewski-Davis, J. Maisano, M. McIntosh, and et al. Best practices for 3d data preservation. In J. Moore, A. Rountrey, and H. Kettler, editors, *3D Data Creation to Curation: Community Standards for 3D Data Preservation*, Chicago, 2022. Association of College and Research Libraries.