/open

RheinlandPfalz
MINISTERIUM FÜR ARBEIT, SOZIALES, TRANSFORMATION UND DIGITALISIERUNG

HANDREICHUNG DATENFORMATE

Der gesetzliche Rahmen für offene Daten gibt vor, dass Informationen in maschinenlesbarer Form bereitgestellt werden sollen. Tim Berners-Lees **5-Sterne-Modell** konkretisiert dies: Es definiert die Verwendung eines strukturierten sowie offenen Formats als Anforderung für Maschinenlesbarkeit und um die Nutzbarkeit der Daten zu erhöhen. Ziel unserer **Handreichung Datenformate** ist es, Datenbereitstellenden praktische Hinweise zu vermitteln, die es ihnen ermöglichen, Informationen als offene Daten zu veröffentlichen, die nach dem 5-Sterne-Modell mindestens eine Bewertung von 2 bzw. 3 Sternen erhalten.

Erklärung zentraler Terminologie

Die Anforderung der **Maschinenlesbarkeit** ist erfüllt, wenn Informationen in einem Format gespeichert werden, das von Computern ohne menschliches Eingreifen gelesen und verarbeitet werden kann.

Werden Informationen in einer Struktur, die eine einfache Verarbeitung und Analyse ermöglicht, organisiert, spricht man von einem **strukturierten Format**.

Ein **offenes Format** liegt vor, wenn die zugrundeliegende Spezifikation zur Speicherung digitaler Daten ohne rechtliche und technische Einschränkungen genutzt werden kann.

Datenformate und -strukturen

Häufig werden Daten in Fachanwendungen erfasst, die Informationen in einem bestimmten Dateiformat über eine API oder einen Export bereitstellen. Im Idealfall wurde bei der Konzeption der Anwendung bereits darauf geachtet, das richtige Dateiformat für die Art der Informationen zu wählen.

Bestenfalls sollte sich die Struktur der Daten bei einer Fachanwendung an einem Standard orientieren, wie zum Beispiel **Schema.org**. In jedem Fall aber sollten Daten, die regelmäßig erfasst werden, stets in der gleichen Struktur erfasst werden. Zusätzlich ist es empfehlenswert, die Vollständigkeit der Daten anzustreben oder Leerfelder explizit als sogenannte **Missings** zu kennzeichnen. Zudem sollte eine sinnvolle Zusammenfassung gleichartiger Datensätze einer Einzelveröffentlichung vorgezogen werden. Weitere Informationen finden Sie in unserer Handreichung zum Thema **Linked-Open-Data** unter open.rlp.de.

/open

Welche Datenformate eignen sich für welche Zwecke?

Häufig liegen Informationen in dem Format vor, in dem sie erstellt wurden. Doch wenn Informationen händisch erstellt werden, kommen dabei oft Anwendungen zum Einsatz, die über umfangreiche Funktionen zur Formatierung verfügen. Denn Formatierungen erleichtern es Menschen, Informationen zu verstehen. Im ungünstigsten Fall ist jedoch das verlustfreie Öffnen solcher Daten nur mit einem proprietären Softwareprogramm möglich. Dadurch wird die Wiederverwendbarkeit der Daten eingeschränkt.

Im Arbeitsalltag sind maschinenlesbare Formate normalerweise weniger übersichtlich für die Bearbeitung durch Menschen, da sie für die Verarbeitung durch Computer konzipiert sind. Im Umkehrschluss erfüllen sie jedoch die oben genannten Kriterien der Strukturiertheit, Maschinenlesbarkeit und Offenheit. Sie eignen sich daher für die Bereitstellung von offenen Daten. Es bietet sich also an, bereits vor der Erfassung der Daten ein geeignetes Dateiformat zu planen.

Folgende Tabelle dient als Orientierungshilfe:

Art der Informationen	Empfohlene Dateiformate	Beispiel
Fließtext	EPUB, XML, JSON	Protokolle
Tabellarische Informationen	CSV (und Abwandlungen wie TSV)	Haushaltspläne
Hierarchische Daten	JSON, XML	Organisationspläne
Geodaten	GeoJSON, KML	Bebauungspläne
Rastergrafiken	JPEG 2000, PNG, TIFF	Fotos
Vektorgrafiken	SVG	Flaggen
Netzwerke	RDF (siehe separate Handreichung)	Linked Open Data

Was unterscheidet diese Dateiformate?

Bei **CSV (Comma-separated values)** handelt es sich um Textdateien, bei denen die verschiedenen Werte eines Eintrags durch Kommata getrennt werden. Die einzelnen Einträge werden durch Zeilenumbrüche getrennt. Bei der Speicherung sollte die Codierung **UTF-8** verwendet werden. Außerdem ist es sinnvoll, dass Leerzeichen vor den Kommata vermieden werden.

JSON (JavaScript Object Notation) speichert Daten in Form von Schlüssel-Wert-Paaren, Arrays und Objekten. Die Struktur von JSON erlaubt beliebig tief verschachtelte Strukturen, da ein Datensatz auch weitere Datensätze enthalten kann. Als Zeichenkodierung kommt standardmäßig UTF-8 zum Einsatz.

GeoJSON nutzt JSON, um Geodaten nach der Simple-Feature-Access-Spezifikation zu beschreiben. Durch das Format lassen sich Geometrien wie Punkte, Punktwolken oder Polygone abbilden.

KML (Keyhole Markup Language) ist eine Auszeichnungssprache, die der Beschreibung von Geodaten dient. Sie nutzt die Syntax von XML.

XML (Extensible Markup Language) ist eine Auszeichnungssprache, die sich durch ihre Menschen- und Maschinenlesbarkeit auszeichnet. Daten werden in einer hierarchisch strukturierten Textdatei abgebildet. Elemente dienen dazu, ein Dokument zu strukturieren. Sie umschließen den Inhalt mit einem Starttag und einem Endtag und können über Attribute mit weiteren Metainformationen angereichert werden. Elemente können Text oder weitere Elemente enthalten.

JPEG 2000 (Joint Photographic Experts Group), **PNG (Portable Network Graphics) und TIFF (Tagged Image File Format)** sind Formate zur Speicherung von Rastergrafiken. Dabei haben die Formate unterschiedliche Anwendungsfälle. PNG wird überwiegend für Grafiken im Web genutzt, während TIFF besonders gut für Bilder und Grafiken geeignet ist, die gedruckt werden sollen. JPEG 2000 ermöglicht die verlustfreie Komprimierung und kann für die Archivierung von Bildern genutzt werden.

SVG (Scalable Vector Graphics) basiert auf XML und dient der Beschreibung von Vektorgrafiken. Es handelt sich um eine empfohlene Spezifikation des World Wide Web Consortiums.

EPUB (electronic publication) dient der Speicherung von E-Books. Das Format setzt dabei auf freie Standards wie XML und XHTML.

/open

Welche Datenformate eignen sich für welche Zwecke?

Geläufige Tabellenkalkulationsprogramme können zur Generierung von Dateien im CSV-Format genutzt werden. Dabei gibt es einige Punkte zu beachten:

- Bei der Speicherung von Informationen unterstützen folgende Empfehlungen die Verarbeitung der Daten durch Computer:
  - „.“ (Punkt) als Dezimalzeichen verwenden (9.5 anstellen von 9,5)
  - Zahlen als absolute Werte ausschreiben (23000000 anstelle von 23 Millionen)
  - Datumsangaben gemäß ISO 8601 (YYYY-MM-DD).
- Tabellenblätter in Einzeldateien aufteilen.
- Zwischenüberschriften, die die Struktur der Tabelle aufbrechen und allein dem menschlichen Verständnis dienen, sind zu vermeiden, da sie die Maschinenlesbarkeit beeinträchtigen.
- Abkürzungen vermeiden.
- Überschriften für Spalten verwenden.
- Bei der Verwendung proprietärer Speicherformate werden Nutzende von der Weiterverwendung der Daten ausgeschlossen.
- Bei einer Transformation in ein offenes Format (zum Beispiel CSV) können Hervorhebungen und Formatierungen verloren gehen.

Warum sollten die Dateiformate DOC, DOCX und PDF nicht verwendet werden?

Im Bürokontext sind DOC, DOCX und PDF häufig genutzte Dateiformate, um Textdokumente zu speichern und auszutauschen. Im Kontext offener Daten sollte auf die Formate DOC und DOCX verzichtet werden, da sie proprietär und nicht maschinenlesbar sind.

PDF-Dokumente können, wenn sie als **tagged** PDF generiert sind, eine Inhaltsstruktur enthalten. Häufig ist dies aber entweder nicht der Fall oder die Umsetzung ist nicht standardisiert genug, um eine uneingeschränkte maschinelle Verarbeitung zu garantieren. Das Dateiformat PDF sollte daher nur in Ausnahmefällen für die Bereitstellung offener Daten verwendet werden.

RheinlandPfalz
MINISTERIUM FÜR ARBEIT,
SOZIALES, TRANSFORMATION
UND DIGITALISIERUNG

Herausgeber

Open-Data-Kompetenzzentrum Rheinland-Pfalz im
Ministerium für Arbeit, Soziales, Transformation
und Digitalisierung des Landes Rheinland-Pfalz

Bauhofstr. 9
55116 Mainz

E-Mail: cc-od@open.rlp.de