OGD Schweiz

Empfehlungen für OGD Formate

Per 09.12.2015

1 Einleitung

Um das Potential von offenen Behördendaten nutzen zu können, müssen diese in Formaten[1] bereitgestellt werden, welche möglichst einfach zu verwenden sind. Unter offenen Behördendaten versteht man Datensätze, die kostenfrei und im Idealfall in maschinenlesbarer Form zur Sekundärnutzung allen Interessierten zur Verfügung stehen[2]. Das vorliegende Dokument definiert Kriterien für die Publikation von Datensätzen im OGD Portal und gibt Empfehlungen zu spezifischen Formaten und Schnittstellen ab.

Die OGD-Strategie Schweiz definiert die drei Ziele „Freigabe von Behördendaten“, „Koordinierte Publikation und Bereitstellung der Behördendaten“ und „Etablierung einer Open-Data-Kultur“ (siehe #1, OGD Strategie Schweiz 2014 – 2018). Für die zu verwendenden Formate werden im Ziel „Freigabe von Behördendaten“ klare Vorgaben gemacht:

> „Der Bund stellt der Öffentlichkeit seine für OGD geeigneten Daten in *maschinenlesbaren* und *offenen Formaten* zur freien Weiterverwendung zur Verfügung.“ (#1, S. 3499)

Diese Kriterien finden ihre Entsprechung im Grundsatz 2 „Offene und wiederverwendbare Behördendaten“:

> „Die Daten werden in *maschinenlesbarer Form* angeboten, verständlich und zweckmässig beschrieben und dauerhaft zur Verfügung gestellt. Es sollen möglichst *offene Formate* angewandt werden.“ (#1, S. 3502)

Im vorliegenden Dokument werden die Kriterien „maschinenlesbar“ und „offen“ genauer beschrieben sowie die Bedeutung von "verständlich und zweckmässig beschrieben" erläutert. Davon abgeleitet werden Empfehlungen zur Publikation von Datensätzen abgegeben.

[1] Der Begriff „Format“ bezieht sich auf das Element **dcat:mediaType** in der Metadatenformat-Definition von **DCAT:Distribution** im Projekt „OGD Schweiz“ (Projektergebnisse)
[2] FAQ Open Government Data, http://www.opendata.admin.ch/de/faq#was-ist-open-government-data

2 Ziel des Dokuments & Zielpublikum

Ziel des Dokuments
Dieses Dokument beschreibt die Kriterien für Formate und Austauschformen, in denen die Daten auf der OGD-Plattform zugänglich sein sollen, und gibt Empfehlungen bezüglich der zu verwendenden Formate ab. Die Liste der empfohlenen Formate ist nicht abschliessend. Es gilt der Grundsatz, dass eine Datenpublikation in irgendeinem Format wünschenswerter ist als keine Publikation.

Zielpublikum
Das Dokument richtet sich an Datenproduzenten und Datenlieferanten bzw. die Institutionen, die Daten als OGD zur Verfügung stellen wollen.

3 Grundlagen & Auswahlkriterien

Üblicherweise wird zur Illustration in der Diskussion um *Open Data* und *Linked Data* auf das Fünf-Sterne-Modell von Tim Berners-Lee verwiesen (siehe #4, Berners Lee, Tim, Linked Data). Gemäss diesem Modell lassen sich *Open Data* in fünf Stufen unterteilen, wobei die ersten drei Stufen im Wesentlichen *Open Data* und die zwei letzten *Linked Open Data* beschreiben. Das Sterne-Modell hat aber Schwächen, so sind darin beispielsweise Webservices oder Application Programming Interfaces (APIs) nur ungenügend abgedeckt.

Die OGD Strategie Schweiz konzentriert sich zurzeit darauf, die Kategorie ★★★ zu erreichen, d.h. die Publikation der Daten in offenen Formaten sicherzustellen.

Datensätze, die nicht der Kategorie ★★★ entsprechen, sollen dennoch auf dem OGD Portal publiziert werden können. Das OGD-Portal wird mit den entsprechenden Datenlieferanten Migrationspläne erarbeiten, um langfristig die Kategorie ★★★ zu erreichen.

★	Datenpublikation mit offener Lizenz in irgendeinem Format
★★	Datenpublikation in einem strukturierten Format (bspw. Excel statt Tabelle in PDF)
★★★	Datenpublikation in einem offenen Format (bspw. CSV statt Excel)
★★★★	Verwendung von eindeutigen Identifikatoren (URI) für Entitäten
★★★★★	Verlinkung der publizierten Daten mit anderen Daten, um Kontext zu schaffen

Tabelle 1: Aus #2, Open Government Data – Grundlagenstudie Schweiz, S. 86; #5, 5 ★ Open Data (Webseite)

Am wichtigsten ist, dass die Datensätze mit offener Lizenz publiziert werden. Für verschiedene Nutzer können hingegen verschiedene Formate angeboten werden, um den spezifischen Anforderungen der Nutzer gerecht zu werden.

Die zentralen Kriterien für die Datenpublikation im OGD-Portal sind Maschinenlesbarkeit und Offenheit (siehe #1 OGD Strategie Schweiz). Diese Kriterien Maschinenlesbarkeit und Offenheit sind im Folgenden rot gekennzeichnet.

1) Maschinenlesbarkeit

Maschinenlesbarkeit	Das Format und die enthaltenen Daten können von einem Programm automatisiert verarbeitet werden.

"Maschinenlesbar ist ein [...] Format, das so strukturiert ist, dass Softwareanwendungen konkrete Daten, einschließlich einzelner Sachverhaltsdarstellungen und deren interner Struktur, leicht identifizieren, erkennen und extrahieren können"[3].

Mit dieser Definition ist beispielsweise eine Textdatei erst dann maschinenlesbar, wenn die darin enthaltenen Daten so strukturiert dargestellt sind, dass sie mit einem passenden Algorithmus (Programm) korrekt interpretiert und verarbeitet werden können. Ein (X)HTML-Text (eine Webseite) ist demnach nicht maschinenlesbar - ein Browser kann die Inhalte nur für Menschen lesbar auf dem Bildschirm darstellen. Erst wenn die Inhalte der Webseite semantisch strukturiert ausgezeichnet werden (z.B. unter Verwendung des Vokabulars schema.org), können die Inhalte von einem Programm als Datenstrukturen erkannt und entsprechend verarbeitet werden. Ausserdem ist bei textbasierten Formaten eine standardisierte Zeichenkodierung notwendig (wenn möglich UTF-8), um auch spezielle Zeichen (z.B. Umlaute) korrekt interpretieren zu können.

Um einen passenden Algorithmus zur Verarbeitung der strukturierten Daten programmieren zu können, ist das maschinenlesbare Format noch nicht ausreichend. Die Daten müssen zusätzlich "verständlich und zweckmässig beschrieben" sein. Falls die Struktur der Daten einem Standard folgt, genügt die Angabe des betreffenden Standards (z.B. schema.org). Andernfalls muss die Beschreibung der Datenstruktur in die Metadaten integriert werden.

2) Offenheit[4]

Offenheit	Die Spezifikation des Formats ist offen gelegt und öffentlich zugänglich.

"Offen ist ein Dateiformat, das plattformunabhängig ist und der Öffentlichkeit ohne Einschränkungen, die der Weiterverwendung von Dokumenten hinderlich wären, zugänglich gemacht wird"[5].

Die Offenheit bzw. die Zugänglichkeit der Spezifikation ist für die lückenlose Interpretierbarkeit der durch das Format kodierten Information entscheidend. Dazu gehört auch, dass der Standard verständlich, überschaubar und eindeutig sein soll; dass dieser Standard lizenzfrei ist und weit verbreitet ist. Dieses Kriterium muss erfüllt sein, damit ein Format als offen gelten kann.

Beide Kriterien Maschinenlesbarkeit und Offenheit weisen mehrere Dimensionen auf, die im Folgenden beschrieben werden. Die nachfolgenden, blau gekennzeichneten Dimensionen, sollen von den Datenlieferanten bei der Auswahl geeigneter Formate, bzw. Austauschformen für ihre Datensätze mitberücksichtigt werden.

[3] RICHTLINIE 2013/37/EU DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 26. Juni 2013 zur Änderung der Richtlinie 2003/98/EG über die Weiterverwendung von Informationen des öffentlichen Sektors (http://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32013L0037&from=DE)

[4] Die Kriterien zur Offenheit stammen aus dem Konzeptbericht Ellipse, siehe #9, Konzeptbericht Ellipse zur Archivierung von Geodaten.

[5] RICHTLINIE 2013/37/EU DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 26. Juni 2013 zur Änderung der Richtlinie 2003/98/EG über die Weiterverwendung von Informationen des öffentlichen Sektors (http://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32013L0037&from=DE)

Standard	Das Format wird durch ein Standardgremium periodisch überprüft und aktualisiert (international, national).

Die Standards von (inter-)nationalen Organisationen sind langlebiger und weniger revisionsanfällig, verleihen einem Format also Stabilität. Ein Standardgremium ist grundsätzlich unabhängiger und glaubwürdiger als eine einzelne Firma, welche die Rechte an einem Format besitzt.

Lizenzfrei	Formate sind nicht durch auferlegte Lizenzen in der Herstellung oder Benutzung limitiert.

Verhinderung einer Herstellerabhängigkeit. Verminderung bzw. Ausschaltung des potentiellen Risikos, dass der Lizenzinhaber jederzeit Änderungen vornehmen und durchsetzen kann.

Verbreitung	Das Format ist weit verbreitet.

Grosse Verbreitung bedeutet eine zum heutigen Zeitpunkt abschätzbare relativ grosse Benutzerzahl wie auch eine gewisse Diversität der Benutzer, eine grosse Anzahl existierender Dateien im entsprechenden Format, Unterstützung des Formats durch viele voneinander unabhängige Applikationen und Programmbibliotheken. Ebenso ist die geografische Verbreitung gemeint.

4 Anwendungsfälle und Formate

Je nach Anwendungsfall können offene Daten ganz unterschiedliche Formate aufweisen. Häufig werden die Daten in Tabellenform vorliegen. Aber auch Bilder, Tonaufnahmen oder Videos können Datensätze oder Bestandteile von Datensätzen darstellen. Entsprechend gross ist die Vielfalt von Formaten, in denen Daten auf dem OGD Portal veröffentlicht werden können.

Die folgende Tabelle listet die wichtigsten, für die Publikation auf dem OGD Portal in Frage kommenden offenen Formate zusammen mit dem Verweis auf den jeweiligen Standard auf.

Anwendungsfälle und Datenformate	**Verweis auf Standard / Quasi-Standard** (wenn verfügbar)
Strukturierte Daten	
Comma Separated Values (CSV)	RFC 4180
JavaScript Object Notation (JSON)	RFC 4627
Extensible Markup Language (XML)	W3C REC
Resource Description Framework (RDF)	W3C REC
Office Open XML Spreadsheet (XLSX)	ISO/IEC 29500
Open Document Spreadsheet (ODS)	ISO/IEC 26300-1:2015
Textdokumente / Berichte	
Text (TXT)	ISO Latin-1 (ISO 8859-1) und ISO Latin-9 (ISO 8859-15) Universal Coded Character Set (UCS) (ISO 10646) US-ASCII (ANSI X3.4-1986, bzw. ISO/IEC 646-US oder ISO/IEC 646:1991-IRV): ISO/IEC 646:1991
Extensible Hypertext Markup Language (XHTML)	W3C REC
Portable Document Format (PDF)	ISO 32000-1:2008 (nur PDF 1.7), ISO 19005 (PDF/A)
Office Open XML Document (DOCX)	ISO/IEC 29500
Open Document Text (ODT)	ISO/IEC 26300-1:2015

Geodaten	
GeoJSON (JSON)	http://geojson.org
KML (XML)	OGC KML
GML (Geography Markup Language)	OGC GML
INTERLIS	http://www.interlis.ch
INTERLIS/GML (gemäss eCH-0118)	eCH-0118
ESRI Shapefile[6]	ESRI Shapefile Technical Description
GeoPackage	OGC GeoPackage
GeoTIFF	http://trac.osgeo.org/geotiff
Bild-/Grafikformate	
TIFF (Tagged Image File Format)	TIFF Revision 6.0
JPEG2000	ISO/IEC 15444-1 :2004
PNG	ISO/IEC 15948:2004
SVG	W3C REC
Audio-/Videodaten	
FLAC	FLAC Format Spezifikation
WebM	Web M Documentation
Ogg Vorbis	Vorbis I specification
MPEG4	ISO/IEC 14496-10 Coding of audio-visual objects -- Part 10: Advanced Video Coding ISO/IEC 14496-3 Coding of audio-visual objects -- Part 3: Audio ISO/IEC 14496-14 Coding of audio-visual objects -- Part 14: MP4 file format ISO/IEC 14496-17 Coding of audio-visual objects -- Part 17: Timed Text subtitle format
Wave[7]	Multimedia Programming Interface and Data Specifications 1.0

Weitere Austauschformen	
WMS (Web Map Service) - nur eine Darstellung der Geodaten	WMS-Beschreibung des OGS (Open Geospatial Consortium)
WFS (Web Feature Service) für Geodaten	OGC WFS
SPARQL (Protokoll und Abfragesprache)	W3C SPARQL
ODATA (Open Data Protocol)	http://www.odata.org

Tabelle 3: Quelle: #2, Open Government Data – Grundlagenstudie Schweiz, S. 88; Eigene Ergänzungen

Gepackte Dateien im ZIP Format sind grundsätzlich erwünscht, sofern der Inhalt in einem geeigneten offenen Format vorliegt. Bei grösseren Datensätzen im CSV oder JSON Format ist eine ZIP-Komprimierung zu empfehlen.

[6] Weit verbreitetes Format, jedoch proprietär (ESRI ArcGIS-Software). Es existieren Programmbibliotheken.
[7] Es existiert kein publizierter Standard für WAVE Dateien. Das WAVE-Format ist eine Implementierung des Ressource Interchange Formats (RIFF). Dieses ist als gemeinsame Publikation von IBM Corporation und Microsoft Corporation, August 1991 freigegeben (http://www-mmsp.ece.mcgill.ca/documents/audioformats/wave/Docs/riffmci.pdf).

5 Empfehlungen

Grundsätzlich sind im OGD-Portal strukturierte Daten zu verwenden. Textdokumente und Berichte dienen der Erläuterung der strukturierten Daten. Es sind Formate zu verwenden, welche für den Inhalt der Daten angemessen sind und die der Benutzer sinnvoll weiterverwenden kann.[8] Es ist erlaubt, den gleichen Datensatz in mehreren Datenformaten auf dem OGD Portal zu publizieren.

- Für **strukturierte Daten** sind „einfache" Formate wie CSV, XML oder JSON komplexeren Formaten wie ODS und XSLX vorzuziehen. Bei regelmässiger Publikation kann die Bereitstellung einer programmierbaren Schnittstelle (API)/eines Webdienstes erheblichen Zusatznutzen stiften.
  Proprietäre, nicht-offene Formate (wie XLS) sind zu vermeiden. Dies auch, weil nicht davon ausgegangen werden kann, dass der Benutzer eine entsprechende Applikation zur Anzeige oder Weiterverarbeitung der Daten besitzt.
- Für **Textdokumente und Berichte** sind (X)HTML, XML, ODT vorzuziehen. Falls eine maschinelle Weiterverarbeitung des Texts angestrebt wird („Natural Language Processing"), ist HTML/XML vorzuziehen. Dokumentationen zu angebotenen Datensätzen (Datenmodelle, Codelisten etc.) können als ODT, PDF oder DOCX publiziert werden. Falls es sich um einen Bericht handelt, der strukturierte Daten interpretiert und darstellt, sollten die zugrunde liegenden strukturierten Daten in angemessenen Formaten ebenfalls publiziert werden. Zu vermeiden sind ältere, proprietäre Formate wie DOC.
- Für **vektorielle Geodaten** wird idealerweise ein modellkonformer Austausch vorgesehen, der mit INTERLIS oder INTERLIS/GML (gemäss eCH-0118) sichergestellt werden kann. Zusätzlich kann für einen vereinfachten Zugang zu den Geodaten ein anderes Format des Kapitels 4 zur Verfügung gestellt werden.
- Für **Raster-Geodaten** wird GeoTIFF empfohlen.

6 Referenzen

#	Titel	Quelle, Link
1	OGD Strategie Schweiz 2014 – 2018	http://www.egovernment.ch/umsetzung/00881/00883/index.html?lang=de
2	Open Government Data – Grundlagenstudie Schweiz	http://www.egovernment.ch/umsetzung/00881/00883/index.html?lang=de
3	G8 Open Data Charter (Englisch)	https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207772/Open_Data_Charter.pdf
4	Berners Lee, Tim, Linked Data	http://www.w3.org/DesignIssues/LinkedData.html
5	5 ★ Open Data (Webseite)	http://5stardata.info/
6	Choosing appropriate formats	https://www.gov.uk/service-manual/user-centred-design/choosing-appropriate-formats.html
7	Konzeptbericht Ellipse – Archivierung von Geodaten	http://www.bar.admin.ch/themen/00876/00939/index.html?lang=de

[8] Dieser Abschnitt wurde stark von der britischen Lösung inspiriert, siehe #6, Choosing appropriate formats

7 Anhang

7.1 Formate in der OGD Praxis

Die folgende Tabelle zeigt die am meisten bereitgestellten Formate in den nationalen OGD-Portalen in den USA, UK, D, A und CH:

	Portal					
Format[9]	**US**	**UK**	**EU**	**DE**	**CH**	**Summe**
html	29'877	1'747	2'335	5'204	11	**39'174**
xml	27'255	338	3'409	438	1	**31'441**
csv	9'302	3'797	10'899	7'177	17	**31'192**
pdf	15'379	1'041	1'527	1'917	16	**19'880**
zip	13'309	193	1'263	265	19	**15'049**
xls	4'463	1'835	6'211	510	2	**13'021**
json	8'418	0	2'567	132	0	**11'117**
wms	5'077	1'453	595	3'811	95	**11'031**
rdf	5'849	278	2'032	5	0	**8'164**
xlsx	165	0	54	5'944	6	**6'169**
Weitere	63'824	292	6'426	7'597	1'738	**79'877**
Gesamtdaten	**182'918**	**10'974**	**37'318**	**33'000**	**1'905**	**266'115**

Tabelle 2: Anzahl Formate in div. Portalen, sortiert nach Gesamtsumme mit Stichtag 23.04.2015

Das Containerformat ZIP wird vor allem im US-Portal verwendet. Betrachtet man nur diese fünf ausgewählten europäischen Portale, tritt an die Stelle von ZIP die deutsche Eigenheit „Kartenviewer“, was im wesentlichen Links auf Darstellungsdienste für Geoinformationen sind.

Ein wichtiges Format in der Schweiz ist das Open Document Spreadsheet (ODS) mit 1'701 Datensätzen, mehrheitlich vom Bundesamt für Statistik.

[9] Wo möglich und sinnvoll wurden verschiedene Formatbezeichnungen zusammengefasst. So kennt beispielsweise das US Portal die Bezeichnungen „xls“, „xlsx“ und „excel“. Das deutsche Portal kennt sowohl die MIME Typen „application/csv“ wie auch die Kurzbezeichnung „csv“.
Im US Portal an zweiter Stelle steht das binäre Format „Originator Data Format“ mit 27‘793 Datensätzen, welches auschliesslich von der National Oceanic and Atmospheric Administration (NOAA) produziert wird. Dieses Format wurde hier nicht berücksichtigt.