Leitfaden zur Berücksichtigung von Dokumenten im Metadatenkatalog von „GovData – das Datenportal für Deutschland“

GovData ist das Datenportal für Deutschland. In GovData sollen die offenen Verwaltungsdaten aus den deutschen Open-Data-Portalen aller föderalen Ebenen zentral auffindbar sein.

Als Open-Data-Portal liegt der Schwerpunkt von GovData in zentralen Bereitstellung von strukturierten Daten. Hier unterscheidet sich das Portal von den Transparenzportalen, die auf Länderebene neben Daten in großem bis überwiegendem Maße (unstrukturierte) Dokumente bereitstellen. Mit diesem Leitfaden sollen Regeln formuliert werden, anhand derer eine Abgrenzung vorgenommen wird, welche Inhalte nach GovData übertragen werden und bei welchen eine Übertragung nicht sinnvoll ist.

Für die Entwicklung des Leitfadens wurden folgende Prämissen zugrunde gelegt:

- Strukturierte Datensätze sind aufgrund ihrer maschinellen Weiterverarbeitungs- und Verknüpfungsmöglichkeiten mit anderen Daten generell von übergreifendem Interesse, auch wenn sie lediglich regionalspezifische Informationen beinhalten
- Unstrukturierte Textdokumente sollen grundsätzlich nur in Verbindung mit passenden strukturierten Datensätzen für die Aufnahme in den Metadatenkatalog von GovData vorgesehen werden.

Ausgangssituation

Das ebenenübergreifende Portal GovData ist das *Daten*portal für Bund, Länder und Gemeinden. Es ist von der Grundstruktur als Portal für offene Daten konzipiert und wird auch entsprechend beworben. Es steht Bund, Ländern und Kommunen zur Verfügung. Die Anforderungen an Open Data sind mehrfach definiert worden. Beispielhaft ist hier die Definition der Sunlight Foundation zu nennen, die einen allgemein anerkannten Kriterienkatalog formuliert hat[1], auf den auch bei GovData Bezug genommen wurde. Noch aktueller ist das 5-Star-Modell von Tim Berners-Lee[2]. Beide Definitionen enthalten die Maschinenlesbarkeit strukturierter Datensätze als wesentliches Kriterium.

GovData erhält von den datenbereitstellenden Portalen ausschließlich Metadaten zu den Datensätzen. Diese Metadaten werden über festgelegte Harvesting-Prozesse von den unterschiedlichen Portalen in Form von RDF-Katalogen zur Verfügung gestellt. Neben reinen

[1] Die zehn Open-Data-Kriterien der Sunlight-Foundation https://www.govdata.de/documents/10156/18448/GovData_Open-Data-Kriterien_der_Sunlight_Foundation.pdf/dca8fea0-8e04-4de0-8531-2bc3e8d4abc0; Fortschreibung der Opendata Guidelines: https://sunlightfoundation.com/opendataguidelines/

[2] 5-Sterne-Modell für Offene Daten, https://5stardata.info/de/

Datenportalen sind auch Portale angeschlossen, die eine große Anzahl von Dokumenten über ihr Portal zur Verfügung stellen. Da diese Dokumente vom Grundsatz her nicht im Datenportal GovData erscheinen sollen, werden die Datenbereitsteller gebeten, bei der Bereitstellung der Daten die Übertragung von nichtstrukturierten Dokumenten nach den folgenden Maßgaben auszuschließen.

Eine Veröffentlichung von Dokumenten sollte nur im Ausnahmefall unter den folgenden **inhaltlichen** Rahmenbedingungen veröffentlicht werden:

- Zur begrifflichen Unterscheidung von Daten und Dokumenten:
  Dokumente sind unstrukturierte, textlastige Dateien, die mittels Textverarbeitungsprogrammen, Reportingtools oder ähnlichem erstellt wurden. Sie sind vorrangig im .doc, .docx, .txt, .rtf oder .odt-Format abgelegt oder in eine .pdf-Datei konvertiert. Im Unterschied dazu sind **Datensätze** strukturierte, maschinenlesbare Dateien, die meist in Listen-, Tabellen oder in Form von Graphen dargestellt werden (typisch .xls, .csv, .rdf).

- Zur Frage, in welchen Fällen Dokumente im Metadatenkatalog von GovData aufgenommen werden sollten:
  Dokumente, die im direkten Zusammenhang mit veröffentlichten Datensätzen stehen, sollten stets im Metadatenkatalog von GovData aufgenommen werden. Hierzu sollte für das Dokument und den Datensatz ein gemeinsamer Metadatensatz vorhanden sein. Beispiel: In einer CSV-Datei wird ein Haushaltsplan bereitgestellt. Dann ist es selbstverständlich sinnvoll, zu dieser Datei auch eine Anleitung (beispielsweise als TXT-Datei) und den Haushaltsplan in der veröffentlichten Form (als PDF-Datei) zur Verfügung zu stellen.

Technisch lässt sich die „Einsortierung“ der inhaltlich den vorbezeichneten Kriterien entsprechenden Dokumente wie folgt darstellen:

In dem RDF-Katalog, den Sie für GovData bereitstellen, dürfen (neben den Metadatensätzen zu den strukturierten Datensätzen) nur die Metadatensätze zu den Dokumenten enthalten sein, die den beschriebenen Kriterien entsprechen.

Davon unberührt bleibt der Umgang mit nichtstrukturierten Dokumenten auf Ebene der jeweils eigenen Portale, wo inhaltliche Gründe oder gesetzliche Vorgaben wie Transparenzgesetze eine Bereitstellung auch von Dokumenten notwendig machen.