How To

Metadaten

Praxistipps zur Qualitätssteigerung von Metadaten

Metadaten – also Informationen über die eigentlichen Daten – sind essenziell, um Datensätze auffindbar, verständlich und nutzbar zu machen. Daher ist es wichtig, dass Open Data Datensätze der Berliner Verwaltung mit möglichst guten, aussagekräftigen Metadaten veröffentlich werden.

Diese Handreichung gibt Empfehlungen zur korrekten Formulierung bzw. Angabe von:

- Titeln
- Beschreibungen
- Geografischen Granularitäten (Raumbezug)
- Veröffentlichenden Stellen
- Schlagwörtern (Tags)

ⓘ

Die Metadatenqualität des Berliner Open Data Portals ist derzeit leider ausbaufähig. Die Empfehlungen dieses Dokuments basieren auf Erkenntnissen des „Werkstattprojekts KI & Metadaten“ der ODIS, bei dem fast 3.000 Einträge des Berliner Open Data Portals auf verschiedene Qualitätsaspekte hin untersucht wurden.

1. Titel des Datensatzes

Der Titel eines Datensatzes ist sein Aushängeschild. Er sollte prägnant, aussagekräftig und auch ohne weiteren Kontext missverständlich sein. Ein gut formulierter Titel sollte klar angeben, um welche Art von Daten es sich handelt. Es kann auch sinnvoll sein, den Zeitraum oder räumlichen Bezug anzugeben.

Das zeichnet einen guten Titel aus:

- **Prägnant und informativ:** Der Titel sollte den Inhalt des Datensatzes sofort verständlich machen.
- **Vermeidung generischer Begriffe:** Wörter wie „Daten“, „Liste“ oder „Suche“ tragen nicht zur Verständlichkeit bei.
- **Keine internen Abkürzungen oder Fachjargon:** Der Titel sollte auch für fachfremde Nutzer:innen verständlich sein.
- **Klarer geografischer Bezug:** Falls die Daten sich auf eine bestimmte Institution oder einen Raum beziehen, sollte dies erkennbar sein. Dies macht insbesondere dann Sinn, wenn gleiche oder ähnliche Daten von anderen Stellen/für andere Räume veröffentlicht werden (das gilt z.B. oft für Veröffentlichungen von Bezirksämtern).
- **Zeitliche Einordnung:** Falls die Daten sich auf einen bestimmten, klar definierten Zeitraum beziehen, sollte dies im Titel erscheinen.

Bei den hier genannten Beispielen handelt es sich um echte Datensatzeinträge die zum Zeitpunkt der Erstellung dieses Handouts im Open Data Portal so veröffentlicht waren. Beachte aber: Es gibt immer mehrere Wege und Möglichkeiten gute Metadaten anzulegen. Die Good Practices sind nicht als einzig wahrer Lösungsweg zu verstehen.

Beispiele für Good Practice für Titel

- *„Energetischer Sanierungsfahrplan bezirklicher Gebäude in Lichtenberg 2024“*
- *„Verkehrszählungen an Hauptverkehrsstraßen in Berlin 2023“*

Negativbeispiele für Titel

- *„Suche nach anerkannten Veranstaltungen“*
- *„Förderungen / Finanzen“*

  > Erklärung: Diese Titel sind ohne weiteren Kontext nicht aussagekräftig. Um was für Veranstaltungen handelt es sich? Von wem oder wofür sind sie anerkannt? Was für Arten von Finanzen und Förderungen sind gemeint? Worauf beziehen sie sich?

2. Beschreibung des Datensatzes

Die Beschreibung soll dem potenziellen Datennutzenden tiefere Einblicke in die Inhalte und den Informationsgehalt geben, um den Datensatz zu verstehen und seine Relevanz (für die Zwecke des Nutzenden) zu bewerten. Sie sollte deshalb alle relevanten Informationen enthalten, um den Kontext der Daten verständlich zu machen.

Das zeichnet eine gute Beschreibung aus:

- **Gut verständlich:** Die Beschreibung sollte in ganzen Sätzen formuliert sein und eine klare, verständliche Sprache verwenden.
- **Interne Abkürzungen oder Fachjargon erklären:** Die Beschreibung sollte auch für fachfremde Nutzer:innen verständlich sein. Es sollte daher möglichst auf Abkürzungen verzichten oder diese sollten kurz erläutern werden.

Folgende Fragen geben eine Richtschnur für einen guten Beschreibungstext vor:

- **Was enthält der Datensatz bzw. wie ist er aufgebaut?** Welche wichtigen Spalten oder Felder enthält die Datei; Beziehen sich die Daten auf ein spezifisches Gebiet (z. B. einen bestimmten Bezirk) oder eine bestimmte Raumeinheit?
- **Warum gibt es diese Daten?** Warum wurden diese Daten erhoben; Gibt es gesetzliche oder verwaltungsbezogene Grundlagen für die Datenerhebung?
- **Warum liegen die Daten so vor, wie sie vorliegen? (Methodik):** Wie wurden die Daten erhoben und von wem; Werden die Daten regelmäßig aktualisiert? Falls ja, in welchem Turnus (z. B. täglich, monatlich, jährlich)? Falls nicht regelmäßig, unter welchen Umständen werden sie aktualisiert; Gibt es spezielle Berechnungen oder Ableitungen in den Daten; Sind bestimmte Werte geschätzt oder aggregiert?
- **Wie gibt es wichtiges bei der Nutzung oder Interpretation der Daten zu wissen? (Datenqualität):** Gibt es bekannte Einschränkungen oder Ungenauigkeiten; Wurden Änderungen am Datensatz vorgenommen oder Fehler korrigiert (nur für aktualisierte Datensätze relevant)

Bitte beachte: Die Hinweise oben geben die Idealkriterien einer guten Beschreibung wieder. Entscheide im Einzelfall, ob alle Kriterien für deinen Datensatz

ausführlich beschrieben werden müssen. Es ist klar, dass Aufwand und Nutzen sich die Waage halten sollten.

Es ist übrigens auch möglich, auf ein Dokument oder eine Website zu verweisen, die weiterführende, detaillierte Informationen zum Datensatz enthält. Die wichtigsten Informationen sollten trotzdem im Beschreibungstext zusammengefasst werden.

Beispiele für Good Practice für Beschreibungen

- *„Gewässerkundliche Messdaten“: „Für Berlin stehen aktuelle und historische Messdaten des Landesmessnetzes für Oberflächengewässer (Flüsse und Seen) und Grundwasser (Grundwasserleiter) tagesaktuell bereit. Zu den verfügbaren Daten gehören hydrologische (Wasserstand, Durchfluss) und hydrogeologische (Grundwasserstand, hydrochemische Analyseergebnisse) Messwerte. Weiterhin stehen verschiedene Qualitätsparameter wie Temperatur, elektrische Leitfähigkeit, pH-Wert, Sauerstoffgehalt etc. aus beiden Bereichen zur Verfügung. Es gibt verschiedene Möglichkeiten für den Datendownload: Die API kann genutzt werden, um Daten abzufragen. Hierfür steht eine Dokumentation bereit. Es können aber auch Daten direkt im CSV-Format über die Website wasserportal.berlin.de heruntergeladen werden. Es stehen 3 Varianten von aggregierten Daten zur Verfügung: 1. Einzelwerte der letzten 12 Monate als arithmetisches Mittel der Messwerte über das Zeitintervall von 15 Minuten. 2. Tageswerte ab Messbeginn als Tagesmittelwerte und ggf. mit Tagesmaximum- und Tagesminimumwerten. 3. Monatswerte ab Messbeginn mit Monatsminimum, -mittelwert, -maximum.“*
- *„Parken im Straßenraum“: „Die Daten beschreiben alle Straßenparkplätze des Landes Berlin. Der Datensatz umfasst alle Parkflächen im öffentlichen Straßenraum innerhalb des Berliner S-Bahnringes sowie ausgewählte, angrenzende Gebiete. Die genaue Methodik zur Erfassung des Datensatzes finden Sie auf der Projektseite des eUVM-Projektes. Die Genauigkeit der Daten beträgt entsprechend der ausgeschriebenen Anforderungen mindestens 95 % (Stand Juli 2023). Aufgrund der Anforderung und der Größe des Datensatzes können vereinzelte Fehler nicht ausgeschlossen werden.“*

Negativbeispiele für Beschreibungen

- *„Förderungen / Finanzen“: „Sammlung von Fördermöglichkeiten“*
- *„ Publikationen des Bezirksamtes Steglitz-Zehlendorf“: „Diese Publikationsdatenbank enthält Publikationen des Bezirksamts Steglitz-Zehlendorf.“*

3. Geografische Granularität

Ein Großteil der Datensätze hat einen Raumbezug, das heißt die Informationen lassen sich einem bestimmten Gebiet oder Ort zuordnen. Das Metadatums-Feld „Geografische Granularität“ gibt an, auf welche Raumeinheiten sich die Werte im Datensatz beziehen. Leider werden hier oft Fehler gemacht. Die korrekte Angabe der Granularität ist wichtig, da sie oft darüber entscheidet, ob ein Datensatz für einen Datennutzer relevant ist.

Ein Beispiel: Eine Person benötigt für ein Projekt Daten auf LOR-Ebene. Diese Person kann durch die „Geografische Granularität“ auf Anhieb erkennen, welche

Daten in dieser Granularität vorliegen und welche Daten dagegen zu grob vorliegen (z.B. Einwohnerzahlen auf Bezirksebene).

Die Geografische Granularität sollte sorgsam geprüft und korrekt angelegt werden. Es muss zwischen einer der vom Datenportal vergebenen Angaben gewählt werden. Die gängigsten davon sind:

- **Berlin** (für Datensätze nur Werte für ganz Berlin, ohne weitere räumliche Untergliederung, enthalten)
- **Bezirk** (für Datensätze die einzelne Werte für jeden Bezirk enthalten)
- **Planungsraum** (für Datensätze die einzelne Werte auf Planungsraumebene enthalten)
- **GPS-Koordinate** (für Datensätze die X und Y-Koordinaten für Standorte enthalten)
- **Hausnummer** (für Datensätze die Adressen für Standorte enthalten)

… und einige weitere.

Beispiele für korrekt angegebene Geografische Granularität

- *„Gender Datenreport Berlin 2022 – Politische Partizipation“: „Berlin“*
  - Erklärung: Es handelt sich um statistische Daten, die für das Land Berlin im gesamten erhoben wurden, daher ist die Angabe „Berlin“ korrekt.
- *„Standorte öffentlicher Toiletten“: „GPS-Koordinaten“*
  - Erklärung: Die Daten verfügen über genaue Standortangaben und enthalten X- und Y-Koordinaten, daher ist die Angabe „GPS-Koordinaten“ richtig angegeben.

Beispiel für fehlerhaft angegebene Geografische Granularität

- „Straßenbaumfällungen in Steglitz-Zehlendorf“: „Bezirk“
  - Erklärung: Die Daten haben adressgenaue Informationen, die Angabe Bezirk ist daher falsch und muss zur korrekten Angabe „Hausnummer“ angepasst werden. Die Angabe, dass die Informationen nur für einen bestimmten Bezirk nämlich „Steglitz-Zehlendorf“ vorliegen, sollte stattdessen unter dem Metadaten-Feld „Geographischer Bezug“ erfolgen.

4. Name der „Veröffentlichenden Stelle“

Die korrekte, formalisierte Angabe der datenhaltenden Stelle ist wichtig, da Nutzende so gezielt nach Datenveröffentlichungen bestimmter Senatsverwaltungen oder Bezirksämter suchen können. Sie sollte daher der offiziellen Bezeichnung der Senatsverwaltung oder des Bezirks entsprechen. Unnötige Anhänge oder Spezifizierungen, z.B. die Angabe von Fachbereichen, sollten vermieden werden. Der Zusatz „Berlin“ kann verwendet werden, ist aber nicht zwingend erforderlich.

Beispiele für eine korrekt angegebene veröffentlichende Stelle

„Bezirksamt Friedrichshain-Kreuzberg“ oder *„Bezirksamt Friedrichshain-Kreuzberg Berlin“*

Beispiele für eine korrekt angegebene veröffentlichende Stelle

- *„Bezirksamt Friedrichshain-Kreuzberg“*

Beispiel für eine fehlerhaft angegebene veröffentlichende Stelle

- *„Bezirksamt Friedrichshain-Kreuzberg von Berlin, Abteilung Bauen, Planen und Facility Management, Serviceeinheit Facility Management, Fachbereich Hochbauservice Energiemanagement, FM-Hoch 4.EWG“*

5. Schlagwörter (Tags)

Weitere Tipps zu Tags: ODIS Berlin – Metadaten Tags (https://odis-berlin.de/ressourcen/0_metadaten_tags/)

Passende Schlagwörter (im englischen „Tags“) helfen Nutzer:innen, relevante Datensätze über die Suchmaske des Open Data Portals zu finden, auch ohne den genauen Datensatztitel zu kennen. Die Schlagwörter ermöglichen es außerdem auch, ähnliche Datensätze miteinander in Verbindung zu setzen.

Die Schlagwörter werden vom Datenbereitstellenden selbst angelegt, es gibt also keine Möglichkeit aus einem bestehenden Pool aus Wörtern auszuwählen. Umso wichtiger ist es, aussagekräftige und ausreichend Schlagwörter in den Metadaten anzulegen. Um die Auffindbarkeit zu verbessern, sollten mindestens fünf (gern mehr) inhaltlich relevante Tags verwendet werden.

Bei der Formulierung von Schlagwörtern sollte folgendes beachtet werden:

- **Ausreichend Tags:** Es sollten mind. 5 Tags angelegt werden, gerne auch mehr.
- **Redundante Tags vermeiden:** Tags wie „Berlin“ oder „Bezirk“ sind nicht nötig, da diese Informationen bereits in den anderen Metadaten enthalten sind.
- **Lieber allgemein statt zu spezifisch:** Tags sollten kurz, prägnant und allgemein sein (z. B. „COVID-19“ statt „COVID-19 Erkrankungen nach Altersgruppe“). Der Titel und Beschreibung enthalten ja bereits detaillierte Informationen.
- **Schreibweisen beachten:** Groß- und Kleinschreibung sowie Sonderzeichen werden in der Suche ignoriert. Es sind also keine doppelten Tags für verschiedene Schreibweisen nötig (z. B. nicht „COVID19“ und „Covid-19“). Mehrere Wörter sollten als separate Tags angelegt werden (z. B. „Vermögen“ und „Finanzen“ statt „Vermögens- und Finanzlage“).
- **Keine internen Abkürzungen oder Fachjargon:** Es sollten nicht nur Fach- oder Verwaltungsbegriffe verwendet werden. Stattdessen sollte die Suchweise von Menschen ohne Fachexpertise berücksichtigt werden. (z. B. „Asylbewerber“ und „Sozialleistungen“ statt „Sozialleistungen nach AsylbLG“).

Beispiel für Good Practice bei Tags

- *„Kostenlose und ermäßigte Angebote im Land Berlin“: [Ermäßigung, Kostenlos, Angebote, Leistung, Fairgnügen, Kultur, Bildung, Sport, Freizeit, Museum, Theater, Teilhabe, Berechtigung, berlinpass, berlin-ticket]*

Negativbeispiel für Tags

- *„Kurse der Berliner Volkshochschulen“: [Dienstleistungen im Bereich Bildung, Volkshochschule]*

OPEN DATA INFORMATIONSSTELLE BERLIN

Die Open Data Informationsstelle wird gefördert von der Senatskanzlei und der Investitionsbank Berlin aus den Mitteln des Landes Berlin.

BERLIN
Der Regierende Bürgermeister von Berlin
Senatskanzlei

TECHNOLOGIE STIFTUNG BERLIN