

# Datenqualität und Veröffentlichungsprozesse

Mission Statement Sub-Arbeitsgruppe *Qualitätssicherung auf Open Data-Portalen* der Cooperation Open Government Data Österreich

Version 1.0 - Autoren: *Johann Höchtl, Axel Polleres, Jürgen Umbrich, Brigitte Lutz*

## Mission Statement

Die Sub-Arbeitsgruppe *Qualitätssicherung von Open Data Portalen* verbessert durch technische Maßnahmen und die Erstellung von Leitfäden zur empfohlenen Praxis die Datenqualität aktuell verfügbarer Datensätze und unterstützt durch organisatorische und technische Maßnahmen den Veröffentlichungsprozess, um in Zukunft höhere Qualitätsniveaus, und somit erhöhte Nutzbarkeit und Nachhaltigkeit von offenen Daten zu erreichen.

Der Erfolg dieser Mission sollen durch folgende Ziele evaluiert werden:

- **Arbeitserleichterung:** Die von der Sub-Arbeitsgruppe vorgeschlagene empfohlene Praxis und sowie begleitende technische Unterstützung wird von der überwiegenden Mehrheit der betroffenen Verwaltungseinheiten als eine Arbeitserleichterung wahrgenommen. Ziel ist hier einerseits, die Arbeitsschritte für DatenanbieterInnen bei Veröffentlichung und Wartung zu optimieren. Längerfristig wäre es auch wünschenswert, erlebte Arbeitserleichterungen durch entsprechende Umfragen unter den Daten Providern zu dokumentieren.
- **Messbare Datenqualität:** Die Qualität von offenen Verwaltungsdaten Österreichs wird von den nationalen Stakeholdern als anerkannt hoch bezeichnet; In internationalen Rankings sind österreichische Verwaltungsdaten im Bereich der Daten-Qualität im vorderen Drittel. Die Definition entsprechender Metriken bzw. Recherche bestehender Metriken und der Vergleich mit anderen Portalen nach diesen Metriken fällt als vorbereitende Maßnahme auch in den Aufgabenbereich der Sub-Arbeitsgruppe *Qualitätssicherung auf Open Data-Portalen*.
- **Vorbildfunktion:** National als auch international werden die Qualitätsmaßnahmen und empfohlenen Metriken zur Datenqualität der Cooperation OGD Österreich als gute Praxis der Veröffentlichung von Verwaltungsdaten herangezogen und von anderen Portalen übernommen.

## Motivation und Problemstellung

Die öffentlichen Verwaltung und privatwirtschaftlich geführte Unternehmen im öffentlichen Besitz sind die einzig legitimen Anbieter öffentlicher Dienstleistungen. Als solche unterliegen sie einem erhöhten Qualitätsanspruch und einer besonders kritischen Kontrolle durch die Öffentlichkeit. Die Benutzbarkeit / Qualität von Open Government Data ist damit einem erhöhten und speziellen Interesse ausgesetzt.

Offene Daten schöpfen dann das volle Potential aus, wenn deren Verwendbarkeit durch die AnwenderInnen gewährleistet ist. Mangelhafte oder fehlerhafte Beschreibungen von Metadaten, fehlende Konsistenz von Daten und Dokumenten, sowie Verfügbarkeitsprobleme bei Services können nachhaltig das Vertrauen der potentiellen AnwenderInnen offener Daten beschädigen.

## Ziele und Nicht-Ziele

Die Sub-Arbeitsgruppe Qualitätssicherung von Open Data Portalen hat es sich zum **Ziel** gesetzt, durch Bereitstellung von organisatorischer guter Praxis, technischen Werkzeugen und beratenden Maßnahmen mittel- bis langfristig die Qualität der bereitgestellten offenen Daten zu bewahren oder messbar zu verbessern. Dieses Ziel soll erreicht werden durch:

1. Sammlung, Auswertung und Bereitstellung von nationalen und internationalen Erfahrungswerten bei der Veröffentlichung offener Daten. Darunter fallen Messmetriken, die den Vergleich von Datenportalen ermöglichen (Benchmarking), Vorgehensmodelle der Veröffentlichung und das Alignment zu bestehenden technischen Konventionen wie Ausarbeitungen des W3C (z.B. CSV on the Web, <http://www.w3.org/2013/05/odbp-charter.html>) oder Initiativen wie [Semic.eu](http://Semic.eu). Bestehende Vorgehensmodelle werden in diese Betrachtungsweise vorrangig miteinbezogen.
2. Evaluierung, Konzeption und Implementierung von technischen Werkzeugen. Technische Maßnahmen und Tools können entlang des gesamten Veröffentlichungsprozesses die Qualität von offenen Daten verbessern, etwa in Form von Software-Werkzeugen für DatenanbieterInnen (die die Veröffentlichung und Aktualisierung von Datensätzen erleichtern) und PortalbetreiberInnen selbst (Monitoring-Werkzeuge).
3. Konkrete Handlungsempfehlungen und Ergebnisse, die datenbereitstellenden Einheiten bei der Verbesserung ihrer Daten und Metadaten unterstützen, sowie Verbreitung dieser Empfehlung etwa in Form von Schulungen.

**Nicht -Ziele** der Sub-Arbeitsgruppe sind

1. die Erstellung von Muss-Kriterien zur Teilnahme an Open Government Data. Die Bereitstellung offener Verwaltungsdaten erfolgt überwiegend freiwillig. Begleitende Maßnahmen sollen diesen Prozess unterstützt und nicht erschweren.
2. Kein Mehraufwand. Mit Hilfe von technischen Werkzeugen werden automatisch die Qualitätskriterien überprüft und der Veröffentlichungsprozess unterstützt.
3. "Neuerfindung des Rades", d.h. die Entwicklung komplett neuer, proprietärer Ansätze, oder Portal-Software (sondern vielmehr die Integration und wiederverwendbare Erweiterung bestehenden Infrastrukturen).

## Einbettung der SUB-AG in das Open Data Big Picture

Die Arbeit dieser Sub-Arbeitsgruppe kann andere Arbeitsgruppen informieren/beraten oder ergänzen. Mit diesen Sub-Arbeitsgruppen sind Überschneidungen zu erwarten:

Sub-AG *Bedarf und Anforderungen zu Linked Open Data*: Die Qualität von Suchergebnissen wird durch Anreicherung mit semantischen Annotationen verbessert (Beispiel: Suchbegriff "Baum" - Ergebnis "Walderhebung 2014"); Die Qualität des Datenaustauschs oder -abgleichs kann durch Zuordnung zu gemeinsamen Thesauri verbessert werden;

Sub-AG *Metadatenkonventionen*: Die Beschreibung von Daten durch Metadaten ist ein offensichtliches Qualitätsmerkmal. Ergebnisse und identifizierte Metriken der Sub-AG "Qualität" können für die Sub-AG "Metadaten" relevant sein.

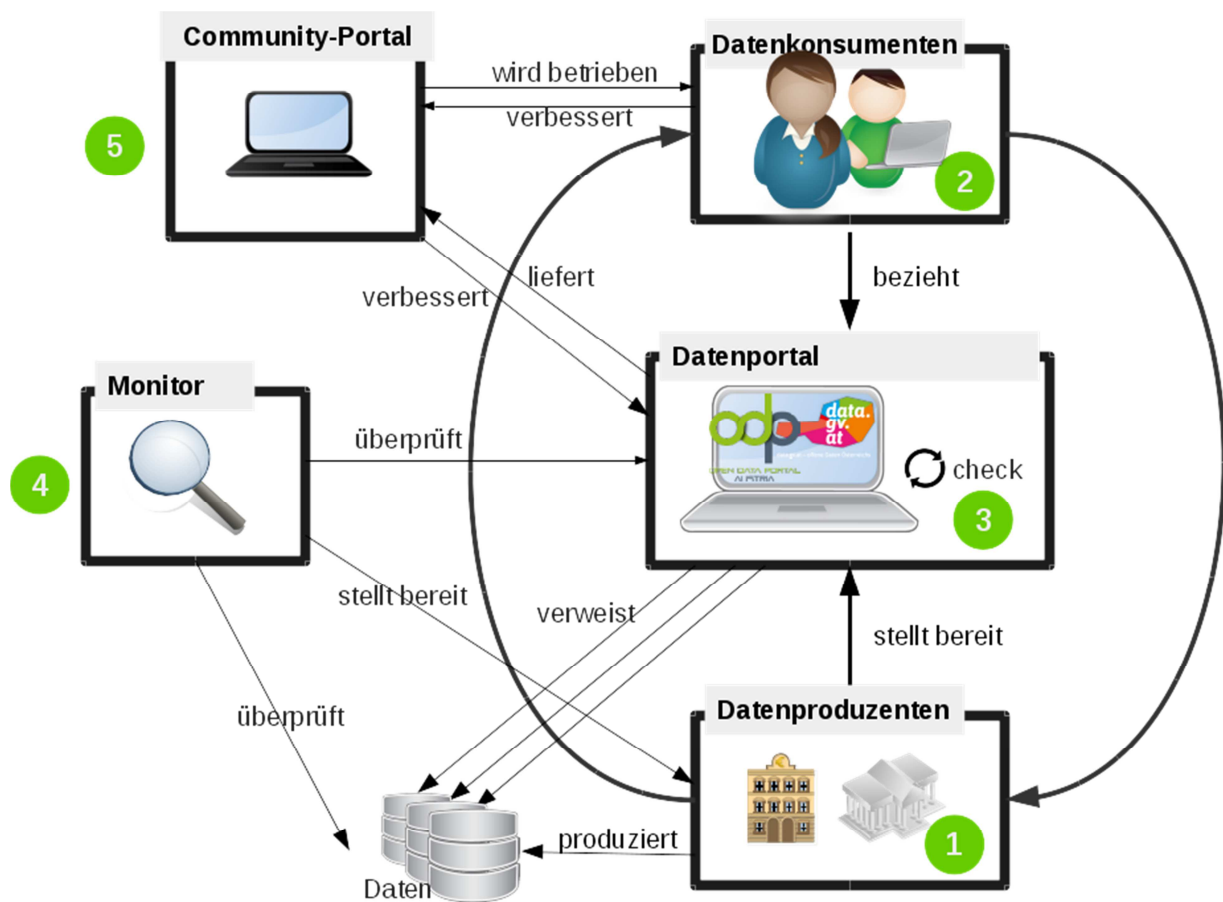
Für die langfristige Etablierung qualitätsvoller Modelle, Vorgehen und Prozesse ist die Verwaltung *gefordert* und die Öffentlichkeit eingeladen, gemeinsam Maßnahmen zu setzen und gute Praxis zu etablieren.

### Qualitätsperspektiven

Ein langfristiges Big Picture zur Wahrung und Steigerung der Qualität von Daten und Metadaten umfasst daher diese Perspektiven und Ebenen:

1. Qualitätsvolle Prozesse und Vorgehensmodelle bei der Erhebung und der Bereitstellung der Daten beim Produzenten
2. Beiträge der Datenkonsumenten
3. Qualitätschecks bei der Eingabe von Daten und Metadaten am Datenportal
4. Monitoring der Datenqualität im Verlauf der Zeit, inhaltliche Überprüfungen
5. Von DatenkonsumentInnen betriebene Datenportale mit benutzerInnengenerierten Inhalten, welche es etwa erlauben, die bestehenden Daten in andere Formate zu konvertieren oder Metadaten anzureichern oder zu verbessern.

Die folgende Grafik bietet einen Überblick zum Open Data Big picture, den Stakeholdern, Informationsflüssen und Angriffspunkten für Qualitätsmaßnahmen:



Für die genannten fünf Angelpunkte der Qualität für offene Daten sind in Österreich bereits konkrete Maßnahmen im Einsatz:

- Das KDZ hat das Open Government Vorgehensmodell bereitgestellt, das den Veröffentlichungsprozess von Daten unterstützt;
- Viele der datenbereitstellenden Verwaltungseinheiten veröffentlichen ihre Daten in einem Community-Prozess. Die Öffentlichkeit kann ihre Wünsche gegenüber den Anbietern formulieren und dabei Anforderungen und Auffälligkeiten kommunizieren.

Weitere Maßnahmen könnten sein:

- Beim manuellen Prozess der Anlage von Metadaten bzw. zusätzlichen Ressourcen können eine Reihe von Maßnahmen den Dateneinpflegenden als Hilfestellung dienen. Mögliche Checks wären: Datumsformat durch Date-Picker, Umfang und inhaltliche Ausdruckskraft der Beschreibung, Angabe der Byte-Größe bei Downloads durch check mit der Ressource, das korrekte Zeichenencoding, usw. ;
- Eine Monitoring-Anwendung kann die tatsächliche Verfügbarkeit von Daten überwachen, zur Datenqualität Auswertungen erstellen und über Reports bzw. Alerts die DatenproduzentInnen auf etwaige Qualitätsprobleme hinweisen;
- Die Öffentlichkeit kann die Daten zu eigenen Zwecken nutzen um beispielsweise ein eigenes Datenportal aufzubauen um funktional mit Kommentarmöglichkeiten,

Verweisen, Hinweisen zu Datenabhängigkeiten und Verwendungszwecken zu erweitern. Diese benutzerinnengenerierten Inhalte verbessern im Rücklauf die eigentlichen Daten der bereitstellenden Einheiten.

## Vorgehensweise und Umsetzung

Mögliche Umsetzungsschritte der Sub-AG "Qualität" können sein:

1. Sammlung nationaler und internationaler guter Erfahrung mit Open Data. Das umfasst Organisationen, Prozesse, technische Implementierungen und bestehende Evaluationskriterien, aus denen gegebenenfalls Maßnahmen abgeleitet werden sollten;
2. Maschinenlesbare Version der österreichischen Metadatenkonvention und davon abgeleitete Formate, die zur Automatisierung von Eingabeprüfungen / Konsistenzüberprüfungen herangezogen werden können ;
3. Überprüfung der Kompatibilität der österreichischen Metadatenkonvention mit DCAT-AP um bei einer Syndizierung österreichischer offener Daten in ein europäisches Datenportal
4. Überprüfung der österreichischen Metadatenkonvention auf ihre allgemeingültige Anwendbarkeit für andere Einsatzdomänen, z.B. als Grundlage für Daten von Privaten oder der Wirtschaft. Möglicher Handlungsbedarf ist hier im Bereich der Klassifikation von Daten gegeben.
5. Implementierung / Erweiterung eines Plugins für CKAN, das während der Dateneingabe Qualitätschecks durchführt. Ein mögliches Plugin zur Implementierung entsprechender Funktionalität wäre <https://github.com/ckan/ckanext-qa>. Checks könnten sein: gültiges Encoding bzw. Encoding vorschlagen, wahre Dateigröße bei download, Struktur der CSV-Datei und Übereinstimmung / Vollständigkeit der Beschreibung;
6. Monitoring der Daten und Metadaten im Zeitverlauf. Sind die Links verfügbar, wie schnell reagieren die Server (Datengeschwindigkeit), welche Inkonsistenzen gibt es in den Daten, wie schnell wurde auf bekannte Fehler reagiert, stimmen die Metadaten noch mit den tatsächlich verfügbaren Daten überein;
7. Verbesserung der Suche auf Datenportalen. Semantische und domänenspezifische Begriffssuche mit der Unterstützung für Synonymen, Hyponymen (relaxed querying), Homonymen. Mit der Verbesserung der Suche kann auch ein Vorschlagsystem für Schlagwörter zu Datensätzen implementiert werden;
8. Mögliche Schaffung eines "Community-Portals", das die Metadaten von data.gv.at und opendataportal.at zusammenfasst und diese weiteren Funktionalitäten anbietet: Diskussion zu Datensätzen, Fehler/Request meldung an Datenverantwortliche mit Tracking. Diese Funktion könnte auch als Bestandteil der bereits bestehenden Portale implementiert werden.

Diese und weitere Maßnahmen werden mit der Cooperation OGD Österreich abgestimmt. Für einige der angeführten Punkte gibt es Implementierungen, die den Stakeholdern der Cooperation bereitgestellt werden können. Für weitere Maßnahmen bedarf es finanzieller Unterstützung die jedenfalls teilweise durch Einwerbung von (Forschungs-)drittmitteln aufgestellt werden sollen.