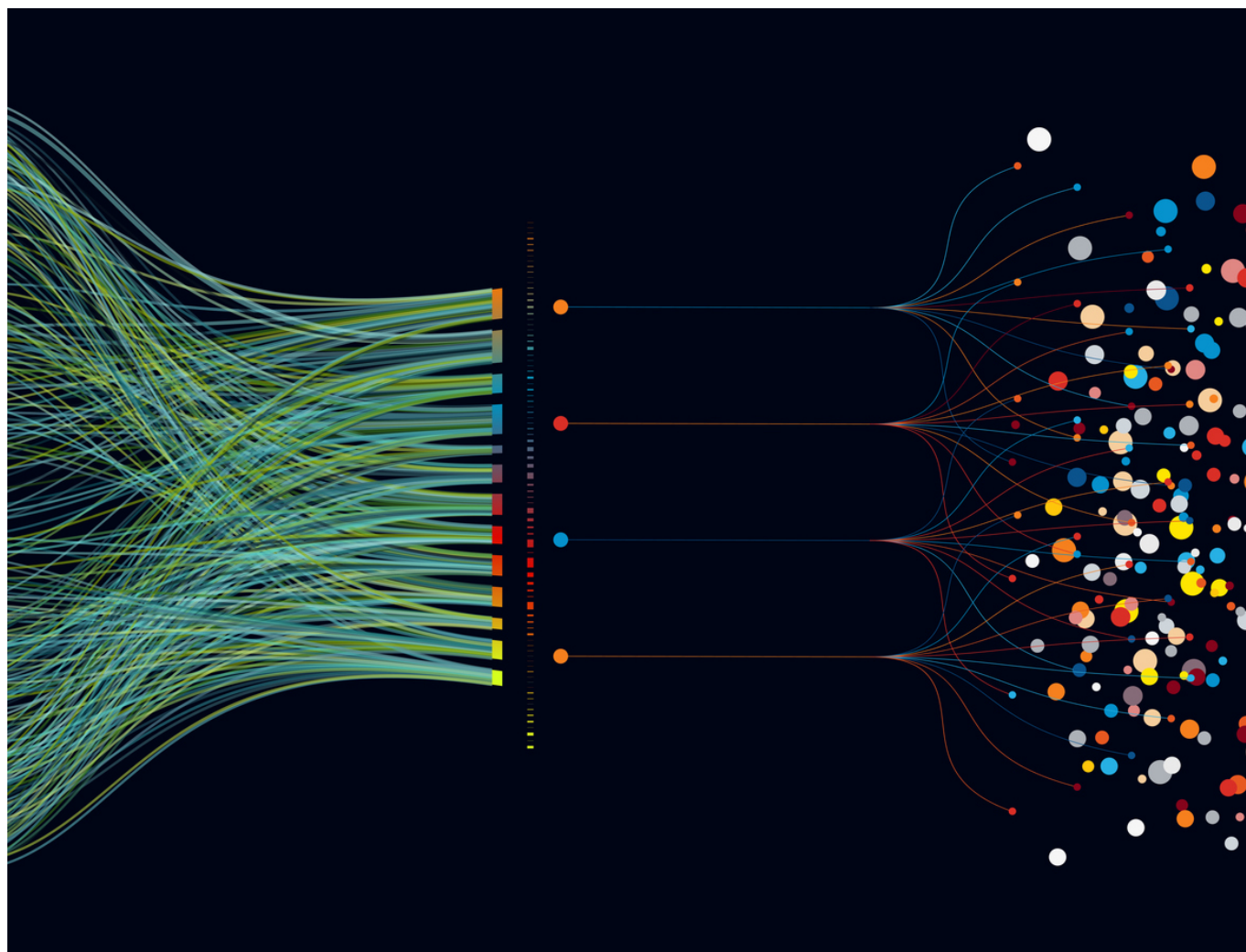


# Die zentrale Rolle von Basisregistern und Normdaten beim Durchbrechen von Datensilos

Autor : Beat Estermann

Datum : 2. Juli 2019



**Basisregister bzw. Normdaten sind zentrale Bestandteile eines Linked-Data-Ökosystems. Zusammen mit gemeinsam verwendeten Datenmodellen bzw. Ontologien sorgen sie dafür, dass Datenbestände auch über Organisationsgrenzen hinweg miteinander verknüpft werden können. Ohne sie wäre "Linked Data" nicht möglich.\***

Im vorliegenden Artikel wird erläutert, was Basisregister bzw. Normdaten sind, und aufgezeigt, weshalb es sinnvoll ist, ihre Publikation als Linked Open Data im Rahmen einer wirksamen Open-Government-Data-Strategie zu priorisieren. Im Kontext eines laufenden Projekts im Auftrag von E-Government-Schweiz wurde hierzu ein Katalog von Priorisierungskriterien entwickelt, der als Grundlage für das Screening bestehender Datenbestände dient und eine zielgerichtete Datenpublikationsstrategie ermöglichen soll.

Linked Data ist die Technologie der Wahl, wenn es darum geht, organisationale Datensilos zu durchbrechen und Datenbestände so bereitzustellen, dass sie möglichst einfach mit Datenbeständen aus anderen Organisationen verknüpft und gemeinsam genutzt werden können. Das Aufbereiten von Daten zu Linked Data ist immer mit einem gewissen Aufwand verbunden, der typischerweise auf Seiten der Datenhalter anfällt. Es ist eine Investition in Interoperabilität zugunsten späterer Datennutzer, zu denen die Datenhalter selber oftmals auch zählen. Das Sicherstellen von Interoperabilität ist keine Einbahnstrasse, sondern bedarf in der Regel des konzertierten Handelns verschiedener Akteure. Im Kontext von Linked Data wird die Interoperabilität durch drei Elemente sichergestellt (Estermann 2018):

- Auf der Ebene der **technischen Infrastruktur** kommen zum Beschreiben und Bereitstellen von Daten Standard-Web-Technologien wie [HTTP](#), [RDF](#) und [URIs](#) zum Einsatz. Sie bilden die Grundvoraussetzung für serverübergreifende, semantische Abfragen der Daten. Dabei werden die Daten typischerweise in [Triple Stores](#) gespeichert, welche über einen [SPARQL](#)-Endpoint abgefragt werden können.
- Die **semantische Infrastruktur** besteht aus Datenmodellen bzw. Ontologien. Diese beschreiben die in den Daten enthaltenen Konzepte und bilden den Zusammenhang zwischen diesen Konzepten untereinander ab. Ontologien gibt es in unterschiedlichen Ausprägungen, die sich vor allem hinsichtlich ihrer Komplexität unterscheiden. Bei Ontologien eines niedrigeren Komplexitätsgrads ist bisweilen auch die Rede von Katalogen, Glossaren, Thesauri oder Taxonomien, die gemeinhin auch unter dem Oberbegriff “kontrollierte Vokabulare” zusammengefasst werden. Werden unterschiedliche Datenbestände mittels derselben Ontologien beschrieben, so sind sie auf der semantischen Ebene interoperabel.
- **Register von Eigennamen**, sogenannte “Named Entities”, dienen dazu, die unterschiedlichen Instanzen einer Klasse eindeutig zu identifizieren. Indem sie für die verschiedenen Instanzen einer Klasse persistente Identifikatoren definieren, ermöglichen sie es, im Rahmen unterschiedlicher Datenbestände Aussagen über dieselbe Person, dieselbe Organisation, dieselbe Verwaltungseinheit, usw. zu machen. Im Kontext von statistischen Ämtern und anderen staatlichen Behörden ist oftmals von “Basisregistern” die Rede. Diese Register haben typischerweise den Anspruch, alle existierenden Instanzen einer Klasse in einer bestimmten Verwaltungseinheit zu verzeichnen, und sind in der Regel dadurch gekennzeichnet, dass eine Behörde einen gesetzlich verankerten Auftrag hat, sie von offizieller Seite her zu führen. Auch in der Definition des European Interoperability Frameworks (Europäische Kommission 2017) wird diese offizielle Pflicht zur Erhebung, Aktualisierung und Bewahrung genannt, betont wird zudem die angestrebte Weiterverwendung der Register durch Dritte. Basisregister können somit als “Masterdaten” der öffentlichen Verwaltung und der Erbringung öffentlicher Dienste angesehen werden. Eine ähnliche Funktion kommt den “Normdateien” zu, wie man sie aus der Bibliothekswelt kennt – sie dienen beispielsweise dazu, Personen oder Werke im Kontext eines Bibliothekskatalogs eindeutig zu identifizieren. Da Linked Data dafür geschaffen wurde, Daten über Organisations- und Domänengrenzen hinweg zu verknüpfen, kommen heute Basisregister und Normdateien oftmals über ihre ursprünglich zgedachte Domäne hinaus zur Anwendung. Insofern unterschiedliche Basisregister oder Normdateien dieselben Instanzen beschreiben, kommen entsprechende Konkordanz-Datenbanken

zum Einsatz – ein prominentes Beispiel aus der Bibliothekswelt ist das [Virtual International Authority File \(VIAF\)](#), welches die Normdateien der Nationalbibliotheken diverser Länder miteinander verknüpft. Ein anderes prominentes Beispiel für einen solchen zentralen Daten-Hub für “Entities” unterschiedlichster Klassen ist [Wikidata](#) (Allison-Cassin & Scott 2018).

Da Basisregister und Normdateien oftmals von Behörden bereitgestellt werden und im Hinblick auf die Etablierung eines gut funktionierenden Linked-Data-Ökosystems eine Schlüsselrolle spielen, stellt die systematische Bereitstellung von Basisregistern und Normdateien einen zentralen Aspekt einer wirksamen Open-Government-Data-Strategie dar. Zu diesem Schluss ist auch die Studie der Berner Fachhochschule zur Weiterentwicklung des Schweizer Datenstandards für Open-Data-Plattformen gekommen (Haller et al. 2018), welche zudem auch auf die wichtige Rolle von Linked Data hinwies, wenn es darum geht, die Auffindbarkeit, die Evaluation und die Nutzbarkeit von offenen Datenbeständen für Dritte nachhaltig zu verbessern.

Um zwei zentralen Handlungsempfehlungen der Studie nachzukommen, gab [E-Government-Schweiz](#) auf Antrag des Schweizerischen Bundesarchivs im Rahmen des Innovationsförderungs-Programms im Frühjahr 2019 ein Projekt in Auftrag, welches zum Ziel hat, die Publikation von Basisregistern und kontrollierten Vokabularen als Linked Open Data systematisch anzugehen. Das Projekt ist dem Ansatz der Aktionsforschung verpflichtet: Einerseits sollen mehrere Basisregister und zentrale Vokabulare auf der Linked-Data-Plattform des Bundes publiziert werden. Andererseits sollen die Grundlagen für eine möglichst effektive und effiziente Datenpublikationsstrategie gelegt werden. Dabei stehen die folgenden Fragestellungen im Vordergrund:

- Welches sind die relevanten Basisregister und zentralen Vokabulare im Hinblick auf die Publikation von LOD durch Schweizer Behörden? Welches sind stichhaltige Kriterien für eine Priorisierung?
- Inwieweit liegen die betreffenden Basisregister und Vokabulare bereits als LOD vor? Welches sind die Gründe, weshalb sie noch nicht als LOD vorliegen?
- Welches sind die praktischen Herausforderungen bei der Aufbereitung von Basisregistern und zentralen Vokabularen zu LOD?

Der Fokus des Projekts liegt auf der Publikation von Linked Open Data durch Schweizer Behörden. Von Interesse sind daher in erster Linie Basisregister und kontrollierte Vokabulare, welche im Zusammenhang mit Schweizer Behördendaten Verwendung finden können. Mit “Basisregistern” sind alle Arten von “Named Entities” gemeint, unabhängig von deren offiziellen Bezeichnung; bei den kontrollierten Vokabularen sind vor allem Taxonomien (z.B. Benennung und hierarchische Struktur von Schweizer Verwaltungseinheiten) und Sets von Ausprägungen für bestimmte Eigenschaften (z.B. Gender) von Interesse. Zur Identifikation entsprechender Daten führt das Projektteam ein erstes Screening von potentiell interessanten Datenbeständen durch und sucht den Austausch mit verschiedenen Schweizer Behörden, die in Sachen Linked-Data-Publikation zu den “First Movers” zählen. In einem weiteren Schritt soll ab dem Sommer 2019 die Schweizer LOD-Community konsultiert und um Inputs im Hinblick auf die Identifikation und Priorisierung von Basisregistern und kontrollierten Vokabularen gebeten werden.

## Priorisierungskriterien

Für die Priorisierung der Basisregister und kontrollierten Vokabulare wurde ein Kriterienkatalog entwickelt, der drei Dimensionen berücksichtigt: (i) Das Nutzungspotenzial, (ii) die (technische und rechtliche) Machbarkeit der Datenpublikation sowie (iii) die Bereitschaft des Datenhalters. Dabei liegt das Augenmerk auf den folgenden Aspekten:

- **Nutzungspotenzial:**

1. Inwieweit sind die Daten mit Datensätzen von Schweizer Behörden verknüpfbar, die bereits als Linked Data publiziert wurden bzw. deren Publikation für den Zeitraum 2019-2020 vorgesehen ist?
2. Wie hoch ist das Nutzungspotenzial im Bereich Open Government Data? – Evaluation anhand konkreter Nutzungsszenarien im Zusammenhang mit offenen Daten.
3. Wie hoch ist das Nutzungspotenzial innerhalb der öffentlichen Verwaltung bzw. innerhalb einzelner Organisationen oder Organisations-Netzwerken? – Evaluation anhand konkreter Nutzungsszenarien im Zusammenhang mit nicht-öffentlichen Daten.

- **Machbarkeit:**

1. Qualität der bestehenden Daten: Welcher Aufwand müsste getrieben werden, um die Daten in ausreichender Qualität bereitzustellen?
2. Vollständigkeit der bestehenden Daten: Welcher Aufwand müsste getrieben werden, um die Daten in ausreichender Vollständigkeit bereitzustellen?
3. Umfang und Komplexität der Daten: Mit welchem Aufwand ist das eigentliche Aufbereiten der Daten zu Linked Data verbunden?
4. Rechtliche Situation der Daten: Dürfen die Daten gemäss aktueller Gesetzeslage freigegeben werden? (Datenschutz, Geheimhaltung, Gebührenordnungen, usw.)
5. Mit welchem Aufwand ist zu rechnen, um die Transformation der Daten zu Linked Data in sinnvollen Zeitintervallen zu gewährleisten? – *B.: Die Daten sollen nicht einmalig publiziert, sondern laufend auf dem aktuellsten Stand gehalten werden. Allerdings können sich Datenbestände hinsichtlich ihrer Aktualisierungsfrequenz sehr stark unterscheiden; von daher variiert auch das technische Dispositiv sehr stark, welches notwendig ist, um eine zeitgerechte Transformation der Daten auf die Dauer sicherzustellen.*

- **Bereitschaft des Datenhalters**

1. Inwieweit ist der Datenhalter **gewillt**, die Transformation der Daten aktiv zu unterstützen bzw. gar dafür die Verantwortung zu übernehmen?
2. Inwieweit ist der Datenhalter **in der Lage**, die Transformation der Daten zu unterstützen oder selber vorzunehmen? In welchem Zeithorizont? Welche Unterstützungsleistungen wären notwendig?

\*Dieser Text ist der 1. Teil eines Artikels, der 2. Teil erscheint in Kürze.

---

## Referenzen

- Allison-Cassin, S., Scott, D. (2018). [Wikidata: a platform for your library's linked open data](#). code{4}lib, Issue 40, 2018-05-04.
- Estermann, B. (2018). [eCH-0205 Linked Open Data](#). Zürich: Verein eCH.
- Europäische Kommission (2017). [Europäischer Interoperabilitätsrahmen – Umsetzungsstrategie](#). Anhang der Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen.
- Haller, S., Estermann, B., Dugga-Winterleitner, A. (2018). [Study in the View of the Further Development of DCAT-AP CH](#). Final Report. Bern University of Applied Sciences.