# Einfacher, schneller, besser – Linked Open Statistical Data

Heute muss man erheblichen Aufwand betreiben, um eine detaillierte Fragestellung mit Daten aus dem Internet zu beantworten. Sucht man zum Beispiel nach der Bevölkerung in der Stadt Zürich und der Stadt Basel, findet man diese Zahlen in einem Excel- oder in einem csv-File im OGD-Katalog. Zuerst muss man aber die jeweiligen Excel-Dokumente suchen und anschliessend darin die gewünschte Kennzahl. Weiter muss man diese beiden Zahlen zusammenfügen, um sie zu vergleichen. Doch was passiert, wenn die beiden Städte unterschiedliche Bevölkerungsdefinitionen verwenden? Sind die Werte dann noch vergleichbar? Wäre es nicht einfacher, mit ein paar Zeilen Code oder sogar einer Suchmaschine diese Datenabfrage zu formulieren und dann auch gleich das Resultat zu erhalten? Und zwar in der Gewissheit, dass beide Städte die gleiche Definition von Bevölkerung verwenden.

## Heutige Datendiffusion

Die Diffusion statistischer Daten konzentriert sich heutzutage auf Excel- und csv-Files. Zudem sind die Daten der statistischen Ämter häufig nicht offen zugänglich, und auch die Formate unterscheiden sich. Die Datenstruktur ist nicht überall gleich, und die Definitionen weichen voneinander ab. Auch auf der Webseite von Statistik Stadt Zürich sind unterschiedliche statistische Daten zur Stadt Zürich als Excel-Tabellen verfügbar. Im Open-Government-Data-Katalog sind weitere detaillierte und gut dokumentierte Daten in maschinenlesbarer Form frei zum Download zugänglich. Diese Daten lassen sich auch mit bescheidenen technischen Kenntnissen auswerten. Sie müssen zur Weiterverarbeitung jedoch heruntergeladen werden und sind für Computer nicht in einer verständlichen Logik verfügbar. Um die einleitende Fragestellung einfacher beantworten zu können, müssen die Daten für Computer mit einer klaren Semantik daherkommen. Dazu setzen wir in Zukunft auf Linked Open Data (LOD) und in unserem Fall auf Linked Open Statistical Data (LOSD).

## Linked Data erklärt

Das Konzept von Internetseiten ist uns allen bekannt. Dokumente, die zum Beispiel Text, Bilder, Videos oder interaktive Grafiken enthalten, werden als HTML-Seiten auf Servern zur Verfügung gestellt. Die Verlinkung einzelner Elemente solcher Webseiten via URL zu anderen Webseiten ist für uns heute selbstverständlich, lässt uns von Information zu Information navigieren und liegt dem Erfolg des Internets zugrunde. Für uns ist dabei die Semantik, also die Bedeutung der Verlinkung oder des Inhalts, verständlich. Wenn wir zum Beispiel nach «Mammut» suchen, können wir gut unterscheiden, ob es sich um die ausgestorbene Elefantengattung oder den Outdoor-Ausrüster handelt. Diese Unterscheidung ist für Computer aber nicht ohne weiteres möglich. Es braucht weitere Informationen, damit auch Computer den Kontext und Inhalt verstehen. Diese fehlende Semantik wird mit LOD in die Daten integriert.

Damit Computer die Semantik verstehen, müssen die Informationen aus strukturierten Daten in ihre Einzelteile zerlegt und danach neu bestückt, als einzelne Information paketiert, im Web publiziert und verlinkt werden. Diese quasi in Atome aufgeteilten Informationspartikel können danach wieder verknüpft und zu «Knowledge Graphs» aufgebaut werden. Durch diese «Knowledge Graphs» können sich sowohl Menschen als auch Computer durchnavigieren, ähnlich wie wir es heute von Webseiten gewohnt sind.

Grundlage ist der W3C-Standard RDF (Resource Description Framework), der Baustein des «Semantic Web», auch bekannt als «Web of Data». Logische Aussagen über beliebige Dinge, sogenannte Ressourcen, können damit formuliert werden.

Jede Aussage besteht aus drei Einheiten (Tripel):
- Subjekt (eine Quelle, die mit einer URI eindeutig identifiziert werden kann)
- Prädikat (eine Spezifikation der Beziehung, die ebenfalls eine URI besitzt)
- Objekt (eine Quelle, mit der das Thema verwandt (URI) ist oder ein Wert)

Das Subjekt und das Objekt stehen also in einer Beziehung, die mit dem Prädikat beschrieben wird. Der Clou ist nun, dass die drei Teile im Web eindeutig über einen Link (URI) verfügbar sind. Die Ausnahme bildet das Objekt, welches auch einfach ein Wert sein kann.
Mit RDF lässt sich somit fast jeder Sachverhalt beschreiben. Es ist klar, dass dafür auch eine geeignete Abfragesprache verwendet werden muss. Bereits 2008 wurde SPARQL vom W3C zum Standard für RDF-Abfragesprachen gemacht. Diese graphenbasierte Abfragesprache erlaubt es uns, das Web der Daten zu durchsuchen und Beziehungen zu entdecken. SPARQL klingt nach einer komplett neuen Sprache, doch wer SQL beherrscht, wird sich leicht mit SPARQL vertraut machen.

**Interview mit Adrian Gschwend (Zazuko GmbH) zu LOD**

Wir werden bei unserer Arbeit mit LOD von der Firma Zazuko GmbH unterstützt. Im folgenden Interview wird Adrian Gschwend die Entwicklung von LOD, die Vor- und Nachteile und mögliche Anwendungen beschreiben.

Seit wann beschäftigen Sie sich mit LOD, und wie kam es dazu?

*Im Jahr 2008 hörte ich einen Vortrag zum Thema RDF, dem Datenmodell hinter Linked Data. Ich fand es damals sehr abstrakt und konnte auf Anhieb nichts damit anfangen. Ein paar Monate später fragte mich ein Kollege um Rat bezüglich eines Kundenprojektes. Der Kunde hatte Daten, die praktisch wöchentlich die Struktur ändern konnten. Uns war klar, dass sich dies nicht mit einer relationalen Datenbank lösen liess. Wir beide dachten aber, das könnte etwas für RDF sein, da das Datenmodell dort ein Graph ist und sich somit von Natur aus deutlich einfacher erweitern und anpassen lässt.*

*Ich befasste mich darauf intensiv mit dem Thema RDF & Linked Data und war begeistert. Endlich ein Datenmodell, welches mit meinen Problemen wachsen kann. Über Umwege gründeten wir 2014 die Firma Zazuko GmbH, welche im LOD-Bereich Consulting und Software anbietet.*

Was sind die Vorteile von LOD?

*Wenn man Daten ablegt, stellt sich immer die Frage, wie man das am besten macht. Seit vierzig Jahren werden dafür mehrheitlich relationale Datenbanken verwendet. Bei solchen muss das Problem von Anfang an gut verstanden werden, damit man sich für die Zukunft nichts verbaut. Mit Formaten wie JSON lassen sich zwar relativ schnell komplexe Strukturen bauen, man verliert aber die Relation zwischen den Daten. Mit Linked Data und RDF kann ich die Vorteile der beiden Welten zusammenbringen: Ich kann mit einfachen Datenstrukturen anfangen und diese Stück für Stück ausbauen und komplexer machen, sofern ich dafür einen guten Grund habe. Das ist deshalb einfach, weil die Datenstruktur von RDF ein Graph ist. Wer sich darunter nichts vorstellen kann, denkt am besten an ein Mindmap oder ein Netzwerk von Freunden: Wenn ich so etwas intuitiv aufzeichne, mache ich daraus automatisch einen Graphen.*

*Bei Linked Data kommt hinzu, dass ich die Daten auf dem Web publiziere und spezifische Informationen zur Verfügung stelle. Ein gutes Beispiel dafür sind Informationen, die uns Google direkt aufbereitet zeigt: Wenn ich bei einer Anfrage für einen Film eine Liste von Aufführungen in Kinos in meiner Nähe zu sehen bekomme, geht dies nur, weil das Kino genau diese Informationen als Linked Data auf seiner Webseite einbettet. So kann die Suchmaschine die genaue Spielzeit des Films lesen und korrekt interpretieren.*

*Und dies ist nur die Spitze des Eisbergs. Linked Data und RDF sind ein offener Standard, der noch viel weitergeht. Die Maschine kann darauf aufbauend Rückschlüsse aus Daten ziehen, die sich mit anderen Technologien nicht oder nur mit viel mehr Aufwand gewinnen lassen.*

Hat LOD auch Nachteile?

*Ich persönlich kann mir kaum mehr vorstellen, ein Problem, das mehr als ein paar Monate überlebt, anders als mit Linked Data zu lösen. In Schulungen und Gesprächen mit Kunden sehe ich natürlich auch, dass viele Entwicklerinnen und Entwickler etwas Mühe haben mit der Datenstruktur, speziell am Anfang. Am Opendata.ch-Event machte ein Teilnehmer im Linked Data Panel eine sehr gute Bemerkung: Linked Data ist am ersten Tag sehr verwirrend, die Datenstruktur erscheint unnötig komplex für einfache Fragestellungen. Ab dem zweiten und dritten Tag stellt man jedoch fest, dass sich komplexere Probleme dafür erstaunlich einfach lösen lassen. Bei klassischen Datenstrukturen ist es wohl mehrheitlich umgekehrt: Am ersten Tag hat man schnelle Erfolge, dafür ist man ab dem zweiten und dritten Tag frustriert, weil komplexere Fragestellungen sehr komplexe Lösungen bedingen. Das trifft es in meinen Augen sehr gut. Ich würde Linked Data, RDF und die Abfragesprache SPARQL nur aufgeben, wenn etwas technisch Besseres kommt. Und da sehe ich nichts Vergleichbares am Horizont.*

*Mit Linked Data gibt es also nicht unbedingt «instant gratification», für viele Entwickler ist dies vielleicht ein Grund, LOD relativ schnell als Lösungsansatz zu verwerfen.*

Wie erklärt man einem Laien, was LOD ist?

*Am besten erklärt man es wohl anhand sozialer Netzwerke: Niemand würde heute noch auf die Idee kommen, Kontakte ausschliesslich in einem klassischen Adressbuch zu speichern. Praktisch alle verwenden irgendeine Art*

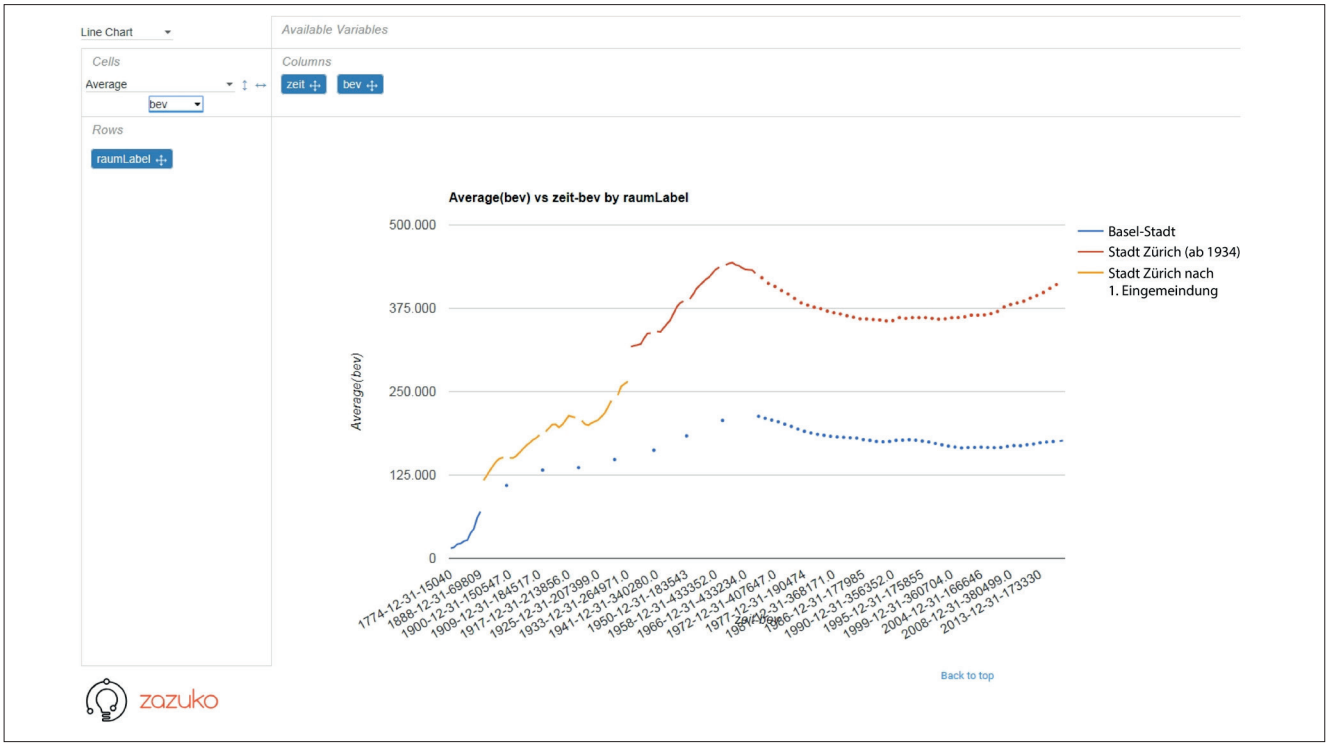
*von sozialen Netzwerken wie Facebook, Twitter, LinkedIn oder Xing. Ich kann so direkt sehen, was andere Leute in meinem Netzwerk machen und was sie beschäftigt. Bei unseren Daten sind wir irgendwie stecken geblieben: Wir haben zwar immer grössere Datensätze, verbinden sie aber nicht oder nur schlecht. Sie stecken in zahlreichen Excel-Dateien, Datenbanken, PDF-Dokumenten und so weiter. Im Firmen- und Behördenumfeld ist dadurch extrem viel wertvolles Wissen nicht für andere zugänglich. Man kann sich Linked Data folglich als soziales Netzwerk für beliebige Daten vorstellen. Natürlich mache ich damit was anderes, aber grundsätzlich kann ich beliebige Daten aus beliebigen Quellen untereinander verknüpfen und Fragen stellen, die sich vorher noch niemand so überlegt hat. Soziale Netzwerke haben komplett neue Interaktionen zwischen Menschen ermöglicht, Linked Data ermöglicht dasselbe für beliebige Daten.*

Was wäre ein typisches Anwendungsbeispiel?
*Mit Linked Data kann ich Fragen beantworten, die man nicht erwartet. Ich denke, das ist für mich der wirklich grosse Anwendungsfall. Datenhalter stellen ihre Daten zur Verfügung und überlassen es den Nutzern, welche Fragestellungen über welche anderen Datensätze sie damit zu beantworten versuchen. Ein Anwendungsbeispiel zu Wikidata wird weiter unten genau beschrieben.*

Wie wird sich LOD weiterentwickeln?
*LOD steht für Linked Open Data, in dem Bereich erwarte ich speziell in der Schweiz mehr offene Daten, welche in RDF veröffentlicht werden. Dies auch im Zusammenhang damit, dass das Bundesarchiv eine Basisinfrastruktur bereitstellt und die Einstiegshürden und Kosten für Datenhalter somit massiv reduziert. Wir sehen aber auch zunehmend Anwendungsfälle, die nicht öffentlich sind. Speziell grössere Firmen kommen immer häufiger zum Schluss, dass in den internen Firmennetzen ein riesiges Datenpotenzial brachliegt und sich Linked Data und RDF dafür eignen, dieses auszuschöpfen. Technisch arbeiten wir und andere Firmen daran, die Einstiegshürden und damit die Komplexität des Stacks zu reduzieren. Gerade am Anfang will man einfach kleine Webapps bauen, welche im Hintergrund auf Linked Data aufbauen. Durch geeignete Werkzeuge und Bibliotheken kann man diese Komplexität reduzieren und die Technologie damit neuen Kunden und Entwicklern zugänglicher machen.*

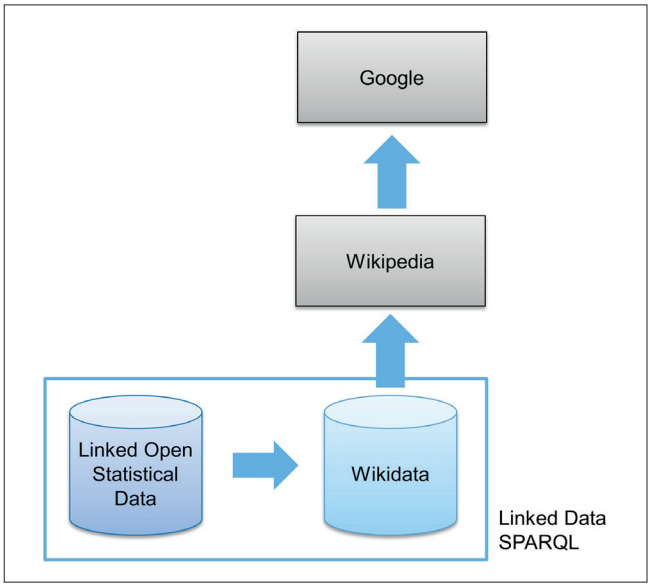
**Anwendungsbeispiele**

Die einleitende Frage lässt sich dank LOSD mit einem Query (Abfrage mit SPARQL) beantworten. Man erhält eine Zeitreihe und sieht auf den ersten Blick die Entwicklung der Bevölkerungszahlen der beiden Städte. Dank



der Semantik der Zahl weiss man, ob die Zahlen auf die gleiche Art definiert sind und eine Verknüpfung sinnvoll ist oder nicht.

Heute kann man nach «Wollishofen Zürich» googeln und findet gleich auf der ersten Suchseite die Einwohnerzahl. Diese kommt von Wikipedia, stammt aber aus dem Jahr 2014. Google und weitere Dienste nutzen Wikipedia als Datengrundlage. Die Wikipedia-Einträge werden häufig manuell gepflegt und sind darum nicht immer auf dem neusten Stand. So kann man auch herausfinden, welche Daten fehlen und ob die Daten übereinstimmen. Man könnte zwar eine einfache Schnittstelle bauen und die Daten so Wikipedia zur Verfügung stellen. Doch LOD kann wie erwähnt mehr. Wikipedia nutzt Wikidata als Datenquelle, und genau diese Datenquelle lässt sich dank LOD einfach und automatisch aktualisieren. LOD liefert die Semantik der Zahl mit. Es wird beschrieben, dass diese Zahl eine bestimmte Art der Zählung der Bevölkerung ist.




In Zukunft stehen alle unsere Daten als LOSD zur Verfügung und können wie hier erklärt via Wikidata auf Wikipedia publiziert werden oder für jegliche Art von Applikationen oder Auswertungen benutzt werden. Wir liefern aber nicht nur die nackte Zahl, sondern auch deren Beschreibung durch Metadaten. Somit wird die Zahl immer gleich interpretiert und lässt sich mit Zahlen gleicher Definition vergleichen.

**Ausblick**

Je mehr Dinge, Ereignisse, Menschen, Orte und natürlich offene Daten im Internet miteinander verbunden sind, desto mächtiger wird der «Knowledge Graph» und damit das «Web of Data». Durch die bessere Zugänglichkeit und die Verknüpfung unterschiedlichster Datenquellen kann neues Wissen aus vorhandenen Fakten einfacher und maschinell unterstützt abgeleitet werden.

Ab Ende August 2018 werden von Statistik Stadt Zürich über 30 Millionen Datentripel im Internet verfügbar, per SPARQL abfragbar und mit anderen Datenquellen verknüpfbar sein. Ende August können die Teilnehmenden der Twist-Hackdays erstmals mit LOSD arbeiten. An zwei Tagen können sie verschiedene Datensätze bearbeiten und die Ergebnisse an den Schweizer Statistiktagen 2018 vorstellen.

Aber keine Angst, falls das alles für Sie noch wie «Rocket Science» klingt. Alle unsere wichtigen Daten können Sie wie bisher auf unserer Webseite finden und von dort beziehen. Zudem arbeiten wir an einem «Statistischen Informationsportal», das die gesamte Linked-Data-Sammlung einfach zugänglich und durchsuchbar macht – auch ohne spezielle Programmierkenntnisse.