UNIVERSITÄT
BERN

Wirtschafts- und
Sozialwissenschaften

Departement
Betriebswirtschaftslehre

Institut für Wirtschaftsinformatik

**Forschungsstelle
Digitale Nachhaltigkeit**

Studie über den Einsatz von Linked Data zur Publikation von Meteodaten

In Zusammenarbeit mit

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

SBB CFF FFS

zazuko

queo

Forschungsstelle Digitale Nachhaltigkeit
Institut für Wirtschaftsinformatik
Universität Bern
Engehaldenstrasse 8
3012 Bern
www.digitale-nachhaltigkeit.unibe.ch

Inhaltsverzeichnis

1 Ausgangslage ... 4

2 Einführung in Linked Data (LD) ... 5

2.1 Das Konzept „Linked Data“ ... 6

2.2 Datenintegration mit Linked Data ... 7

2.3 Technische Anforderungen an Linked Data ... 8

3 Studie ... 12

3.1 Linked Data Technologie ... 12

3.1.1 Technische Prinzipien für die Publikation und Nutzung von Linked Data ... 12

3.1.1.1 Publikation Linked Data ... 12

3.1.1.2 Voraussetzungen ... 13

3.1.1.3 Überführen von MeteoSchweiz Daten nach RDF ... 13

3.1.1.4 Consume ... 14

3.1.1.5 SPARQL ... 15

3.1.1.6 HTTP Resolving / Dereferencing ... 15

3.1.1.7 Linked Data APIs (API Middleware) ... 16

3.1.1.8 Human Layer ... 17

3.1.2 Abgrenzung zu anderen Konzepten ... 18

3.2 Überlegungen zu Sizing / Betrieb ... 18

3.3 Veränderung der Kundenbeziehung- und Bedürfnisse ... 21

3.4 Wissensbedarf & Maturitätslevels ... 27

3.5 Erkenntnisse aus dem Proof of Concept (Show Case) ... 28

3.5.1 Look and Feel der Webapplikation ... 30

3.5.2 Authentifizierung open/closed Data ... 31

4 Weiterführende Informationen ... 33

4.1 Vorhandene Vokabulare im Bereich Meteorologie ... 33

4.2 Bekannte Vorhaben im Umfeld von Linked Data ... 34

4.2.1 Schweizer LD Projekte ... 34

4.2.2 Internationale Projekte ... 35

4.3 Weiterführende Literatur ... 39

5 Anhang 40

5.1 Glossar und Abkürzungen 40

5.2 Verzeichnisse 44

1 Ausgangslage

Die Studie erklärt einerseits die Konzepte von Linked Data zeigt deren Einsatzmöglichkeiten und auch Grenzen für die Publikation von Meteodaten auf.

Die Studie ist in Zusammenarbeit der folgenden Partner entstanden:

Fachpartner und Datenlieferanten:

- Bundesamt für Meteorologie und Klimatologie, MeteoSchweiz
  www.meteoschweiz.admin.ch
- Schweizerisches Bundesarchiv, BAR

  www.bar.admin.ch
- Schweizerische Bundesbahnen SBB
  www.sbb.ch

Umsetzungs- und Technologiepartner:

- Universität Bern, Forschungsstelle Digitale Nachhaltigkeit
  www.digitale-nachhaltigkeit.unibe.ch
- queo swiss, Solothurn
  www.queo.swiss
- Zazuko GmbH, Biel/Bienne
  www.zazuko.com

2 Einführung in Linked Data (LD)

Mit dem Aufkommen von Open Data Prinzipien und entsprechenden Portalen haben viele Behörden begonnen, Datensätze der Öffentlichkeit zugänglich zu machen. Diese Daten müssen heute oftmals vom Datennutzer nachbearbeitet werden, bevor sie in entsprechenden Applikationen mehrwertbringend verwendet werden können.

Linked Data propagiert mit dem Datenmodell RDF und dem dazugehörigen Semantic Web Stack die standardisierte Basis einer Technologie, welche die Maschinenlesbarkeit der Daten erhöht und damit die Wiederverwendung derselben in neuen Anwendungsfällen stark vereinfacht.

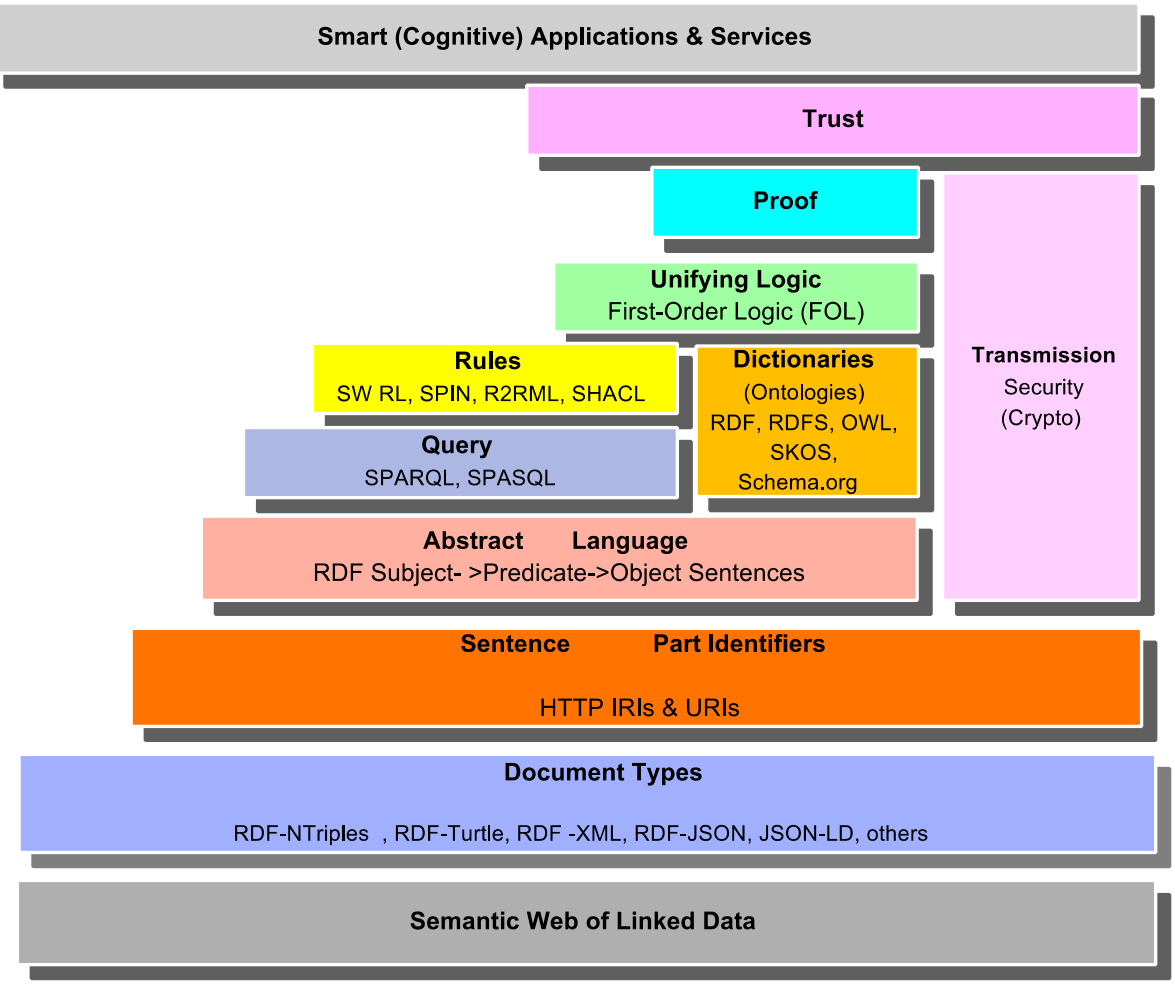


Abbildung 1: Semantic Web Layer Cake

Linked Data baut dabei auf bestehenden Web Technologien auf und wird vom W3C standardisiert, dies ermöglicht einen kompetitiven Markt mit verschiedenen Implementationen der jeweiligen Standards, was wiederum Unabhängigkeit von Herstellern sicherstellt. Die Interoperabilität der Daten und Standards ist dabei Teil des Konzepts, föderale Datenhaltung, Mehrsprachigkeit und Siloübergreifende Abfragen sind Kern des Semantic Web Stacks.

Linked Data ermöglicht Probleme anzugehen, die mit bestehenden Technologien nur schwer oder mit deutlich mehr Aufwand umsetzbar sind. Die Flexibilität des Stacks und dem Graph-basierten

Datenmodell RDF benötigt initial eine gewisse Einarbeitungszeit, sowohl auf Seiten des Datenhalters wie des Datennutzers. Sobald dieser Punkt erreicht ist, kann aber deutlich mehr Wissen wiederverwendet werden, als dies typischerweise in geschlossenen Systemen der Fall ist. Die Komplexität kann dabei kontinuierlich gesteigert werden, es ist möglich mit einfachsten Mittel anzufangen und später mehr und mehr maschinell erledigen zu lassen.

Linked Data ermöglicht, Daten über technische und organisatorische Grenzen hinweg zu verknüpfen. Im Folgenden werden das Konzept „Linked Data“, die technischen Anforderungen an Linked Data und der Aufbau des „Web of Data“ von der einfachen Publikation bis zur Verlinkung von Daten erklärt.

2.1 Das Konzept „Linked Data“

Schon in der Antike entstand die Idee, alles Wissen der Welt an einem Ort – der Bibliothek – zu sammeln, um dessen effiziente und vernetzte Verarbeitung sicherzustellen. Vannevar Bush[1] formulierte diese Vision erstmals auf einer informationstechnischen Ebene und präsentierte, in seinem wegweisenden Artikel aus dem Jahre 1945 den Memex, eine integrierte und vernetzte Wissensmaschine. Aufgrund dieser Vision baute Douglas Engelbart[2], der Erfinder der Computer-Maus, Ende der 60er-Jahre ein erstes Hypertextsystem[3]. Kombiniert mit der Idee der Vernetzung mehrerer solcher Systeme über das Internet erfand Tim Berners-Lee 1989 das World Wide Web[4]. Dieses ermöglicht es Menschen überall auf der Welt, erstens Texte und Medien zu publizieren und zu verlinken – das heisst, Beziehungen zwischen den publizierten Elementen herzustellen – und zweitens in diesen Publikationen Nachforschungen anzustellen.

Die folgenden technischen Grundlagen haben das Wachstum des World Wide Web ermöglicht:

1. Offene und frei zugängliche Standards für Publikation und Zugänglichkeit von Publikationen sowie deren Verknüpfung über sog. Links.
2. Offene Autorenschaft, d.h. im World Wide Web kann jedermann publizieren.
3. Freies Referenzieren anderer Publikationen, d.h. jeder kann sich auf andere Inhalte im World Wide Web beziehen.

Die Inhalte des World Wide Web sind darauf ausgerichtet, durch Menschen konsumiert und interpretiert zu werden. Diese Inhalte lassen sich jedoch nicht automatisch kombinieren, aggregieren und zu neuen Aussagen verdichten, da sie nicht maschinenlesbar respektive maschineninterpretierbar sind. Die automatische Verarbeitung durch Computer benötigt zusätzliche Informationen über die Gestaltung und Form der Daten. Diese Art von Informationen werden Metadaten, also Daten über Daten, genannt. Während frühere Ansätze für die automatische Verknüpfung von Daten aus verschiedenen Quellen die aufwändige Erstellung von globalen Metadatensystemen voraussetzten, proklamieren neuere Ansätze ein Datenformat, welches die Verknüpfung von Daten einfacher ermöglicht – das Konzept direkt miteinander verknüpfter Daten, „Linked Data“.

[1] Vannevar Bush, As we may think. In: Atlantic Monthly 176, S. 101-108. Siehe http://www.theatlantic.com/magazine/archive/1969/12/as-we-may-think/3881/.

[2] Engelbart, D. & English, W. (1968). A research center for augmenting human intellect. AFIPS Fall Joint Computer Conference. p. 295-410 Siehe http://www.dougengelbart.org/pubs/augment-3954.html.

[3] Hypertext/-media Systeme sind Systeme, die Sammlungen von Texten und/oder Medien nicht nur zur Verfügung stellen, sondern deren Daten auch intern miteinander verlinken.

[4] Berners-Lee, Tim, Weaving the Web. Siehe https://www.w3.org/People/Berners-Lee/Weaving/Overview.html

Linked Data benötigt insbesondere keine vorgängige teure Abstimmung über global gültige Metadaten. Im Gegensatz zur herkömmlichen Datenintegration muss also nicht vorher ein global akzeptiertes Schema entwickelt werden.

Zusätzlich haben Linked Data die folgenden wesentlichen Vorteile:

Linked Data liefert bestmögliche Antworten auf Anfragen

Linked Data zielt aufgrund des genutzten Datenformats auf die bestmöglichen Antworten ab, welche sich nicht in Links auf Dokumente oder Webseiten erschöpfen, sondern die Information direkt vermitteln. So ergibt – um ein einfaches Beispiel zu verwenden – eine Suche nach der Bevölkerungszahl der Schweiz nicht eine Liste von Webseiten oder Dokumenten mit dieser Information, sondern direkt die gewünschte Zahl.

Linked Data ermöglicht pay-as-you-go Datenintegration

Linked Data benötigt keine teure Vorinvestition in die Integration und Interoperation der verschiedenen Datenquellen. Durch das standardisierte Datenmodell können zusätzliche Daten jederzeit in einem bestehenden Software-Stack eingebunden werden. Das Graph-basierte Datenmodell lässt sich anpassen, ohne das teure Schemaanpassungen notwendig sind. RDF agiert dabei als agiles Datenmodell, welches mit der Komplexität des Problems wachsen kann.

Linked Data ermöglicht die inhaltliche Anreicherung der Daten

Linked Data kann durch Verknüpfung mit weiteren Daten aus der Linked Data Cloud inhaltlich angereichert werden („mashups“). So können zum Beispiel für die Stadt Bern Linked Data aus der Statistik und zusätzlich aus DBpedia (Wikipedia aufbereitet als Linked Data) zu einem Porträt aus Zahlen, Bildern und Text verbunden werden.

Diese Vorteile ermöglichen es Datenpublizisten, ihre Daten ohne allzu grosse Rücksicht auf andere Daten zu veröffentlichen. Solange die Daten genügend beschrieben sind, wird es Wege geben, diese mit anderen ebenso beschriebenen Daten zu verknüpfen und weiterzuverwenden.

2.2 Datenintegration mit Linked Data

In der Vergangenheit wurde bei der Datenpublikation immer wieder die teure Datenintegration zum Stolperstein. Tatsächlich müssen für die Integration von zwei Datenquellen sowohl die Semantik als auch die Syntax aufeinander abgestimmt werden. So muss man sich, zum Beispiel, bei Koordinaten aus zwei Quellen darauf verständigen, um welche Koordinaten es sich handelt (Landeskoordinaten[5] oder Globale Koordinaten) und darauf, in welcher Form die verschiedenen Werte aufgeschrieben werden (nur in Grad mit Kommastellen oder in Grad, Minuten und Sekunden). Eine vorherige Abstimmung dieser Datenquellen würde ein globales Datenmodell voraussetzen – eine Zielsetzung, welche schon innerhalb kleinster Organisationen oft unerreichbar ist, da viel zu aufwändig und global als unmöglich betrachtet wird.

Deshalb setzten Linked Data auf eine spätere Datenintegration, bei welcher die Benutzer der Daten entscheiden, wieweit die Integration gehen soll. Bei vielen Analysen im Open Data Bereich reichen

[5] Selbst bei den gängigen Landeskoordinaten gibt es verschiedene Standards (siehe http://de.wikipedia.org/wiki/Schweizer_Landeskoordinaten).

solche Ad-hoc-Integrationen, oder es genügen statistische Genauigkeiten, wie uns Google eindrücklich zeigt[6].

2.3 Technische Anforderungen an Linked Data

Es stellt sich nun die Frage, welche Voraussetzungen zur Erstellung von Linked Data notwendig sind. Insbesondere stellt sich für Datenpublizisten das Problem, welche technischen Rahmenbedingungen für die Veröffentlichung der Daten eingehalten werden müssen. Diese Thematik wurde im letzten Jahrzehnt unter dem Begriff „Semantic Web“[7] erforscht. Aufbauend auf den Prinzipien des World Wide Web überlegte man sich, welche technischen Bausteine ein weltumspannendes „Web of Data“ ermöglichen. Daraus ergaben sich die folgenden Anforderungen:

Unter Nutzung und Einhaltung der bekannten und geltenden Prinzipien des World Wide Web kann das weltumspannende „Web of Data“ entstehen, wenn,

erstens, alle Daten und Metadaten publiziert werden,

zweitens, ex ante keine Datenintegration zur Publikation notwendig ist,

drittens durch diese Art der Datenpublikation ein System aufgebaut werden kann, welches bestmögliche Antworten liefert und welches,

viertens, eine einfache Integration weiterer Daten ermöglicht.

Die Semantic Web Gemeinschaft ist sich bewusst, dass nicht alle Datenpublizisten per sofort alle Anforderungen an ein „Web of Data“ erfüllen können. Die mit der Datenaufbereitung verbundenen Aufwände würden dies verunmöglichen. Deshalb hat Tim Berners-Lee das folgende Fünf Sterne Modell für den schrittweisen Aufbau des „Web of Data“ proklamiert.[8]

[6] Siehe dazu Freebase (http://www.freebase.com/; eine von Google betriebene offene Datenbank), Google Fusi-on Tables (http://www.google.com/fusiontables/Home/; eine von Google betriebene Datenintegrationsapplika-tion) sowie Google-refine (http://code.google.com/p/google-refine/; eine Datenreinigungsanwendung von Google).

[7] siehe http://swsa.semanticweb.org/content/international-semantic-web-conference-iswc. Eine der ersten Publikationen, welche die Vision beschreibt ist: Tim Berners-Lee, James Hendler & Ora Lassila: "The Semantic Web", Scientific American 284(5):34-43 (May 2001).

[8] Siehe http://www.youtube.com/watch?v=ga1aSJXCFe0 und http://inkdroid.org/journal/2010/06/04/the-5-stars-of-open-linked-data/ für eine Erklärung.

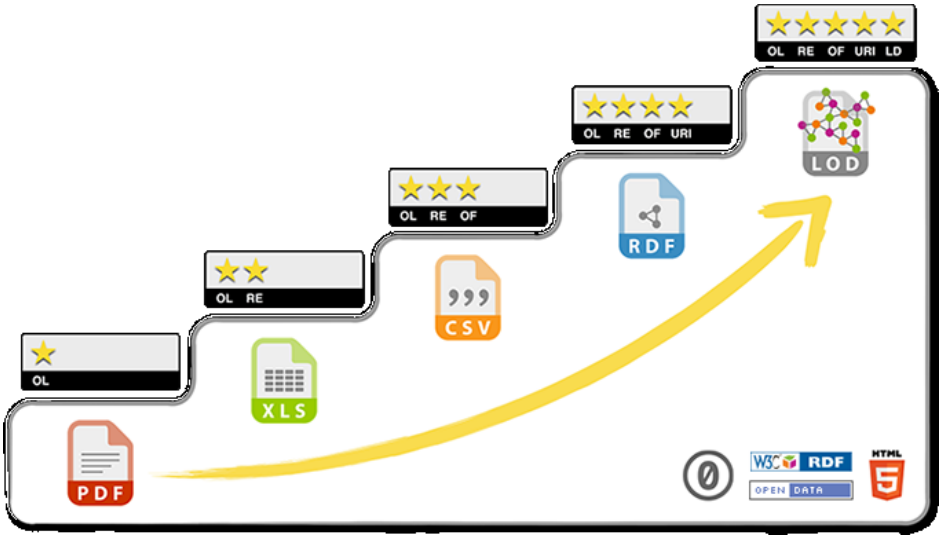


Abbildung 2: Das Fünf Sterne Modell des Datenpublizierens [9]

Sterne	Stufe	Beschreibung
*	1	Publiziere Daten in irgendeinem Format, z.B. PDF-File.
* *	2	Publiziere die Daten in einem strukturierten Format also z.B. als maschinenlesbares Excel-File, statt als Bild eine Tabelle.
* * *	3	Publiziere die Daten in einem nicht-proprietären, offenen Format also z.B. als CSV- statt als Excel-File.
* * * *	4	Verwende eindeutige Identifikationen für Entitäten, damit andere auf deine Daten verweisen können. Zur Erfüllung dieser Anforderungen entwickelte die internationale Forschungsgemeinschaft rund um das Semantic Web ein spezielles Datenformat: das Resource Description Framework (RDF)[10]. RDF geht von der Erkenntnis aus, dass sich die Beziehungen zwischen Daten als Graphen modellieren lassen. Die Knoten repräsentieren dabei die Entitäten und die Kanten die Beziehungen zwischen den Entitäten. Beschrieben werden die Daten in Tripeln (Subjekt – Prädikat – Objekt, z.B. Bern – hat Einwohner – 139 089). RDF ist ein offenes Datenformat, in dem Entitäten durch eine globale Identifikation (URI) und durch Beziehungen zu anderen Entitäten oder Zeichenketten spezifiziert werden. RDF ist sehr flexibel und erlaubt die Publikation von Metadaten in demselben Format. Zusätzlich erlaubt diese Spezifikation die Verteilung der Daten auf verschiedene Publikationen. RDF erfüllt somit alle Anforderungen an ein „Web of Data“.
* * * * *	5	Verlinke die publizierten Daten mit anderen Daten zur Schaffung von Kontext. Das Format unterstützt die dezentrale Verlinkung von Entitäten. Wenn die Publikation A zum Beispiel die Stadt Bern beschreibt, dann kann eine Publikation B mit Untersuchungsgegenstand Schweiz bei der Erwähnung der Hauptstadt einfach auf die entsprechende Stelle in der

[9] http://5stardata.info/en/

[10] http://www.w3.org/RDF/.

Sterne	Stufe	Beschreibung
		Publikation A verweisen. Es ist hierbei zu beachten, dass die Erfüllung dieses Zieles etwas komplizierter ist als im traditionellen Web, da die Links genau auf eine Entitätsbeschreibung in einer anderen Publikationen verweisen müssen, und nicht nur generell auf das Dokument.

Tabelle 1: Fünf Sterne Modell

Das Fünf-Sterne-Modell ermöglicht einen stufenweisen Einstieg in das „Web of Data“, von der einfachen Publikation bis zur Verlinkung der Daten. Jede Stufe stellt dabei eine messbare Verbesserung bezüglich maschinenlesbarer und verknüpfter Nutzung dar. Anzumerken ist hierbei, dass die Verwendung von RDF für die ersten drei Stufen kein Thema ist. Die Publikation auf den Stufen 4 und 5 werden durch RDF allerdings dermassen erleichtert, dass keine ernst zu nehmenden Datenpublikationen auf diesen beiden Stufen bekannt sind, die auf RDF verzichten. Beispiele dafür finden sich im Kapitel 4.2.2.

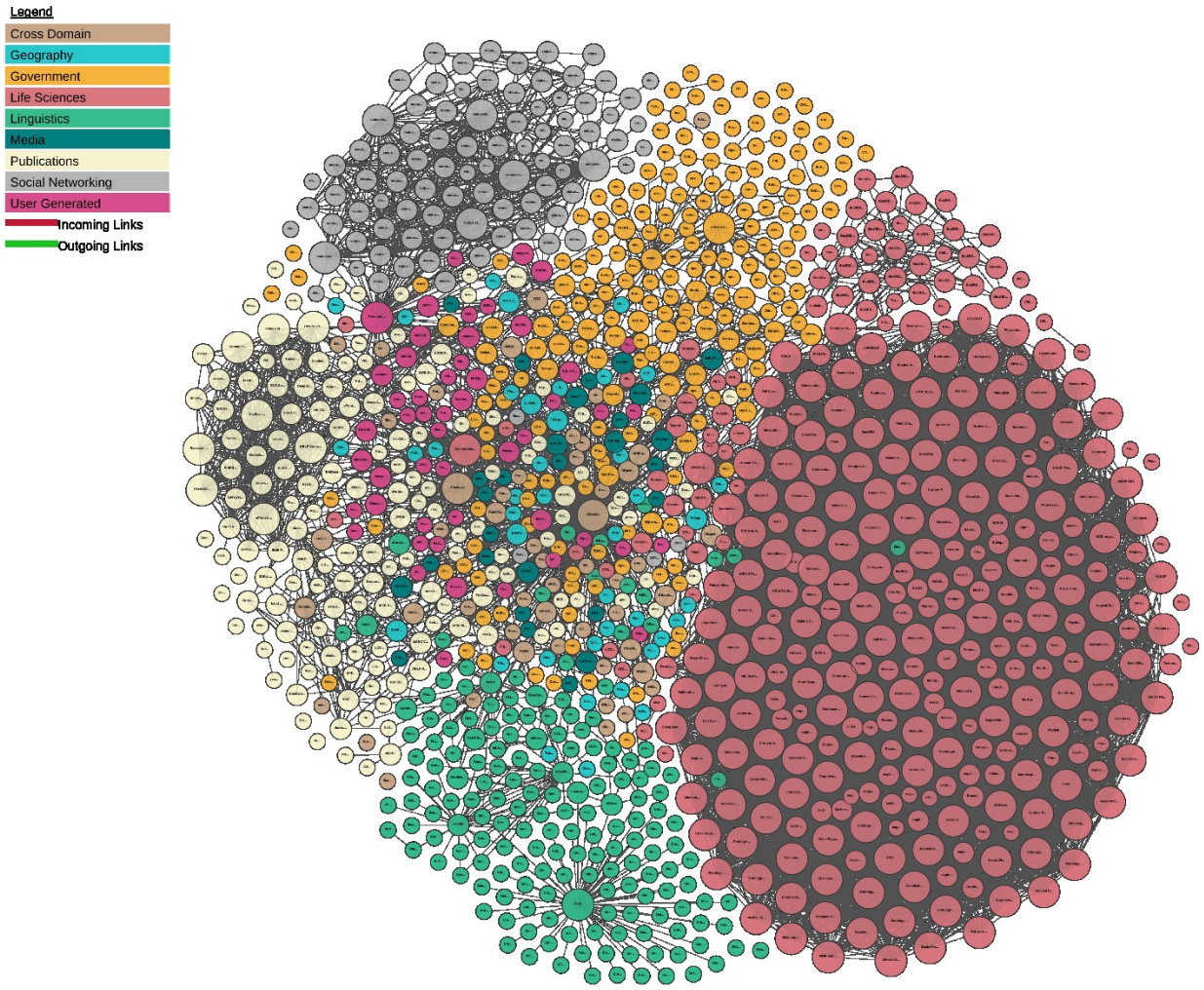


Abbildung 3: Das „Web of Data" – die Linked Data Cloud[11]. Eine interaktive Version findet sich auf http://lod-cloud.net/

Logische Folge ist es, komplexe Abfragen in als Graphen vorliegende Datenmengen in RDF-basierten Datenbanken (Triple Stores) zu ermöglichen. Dazu wurde in den letzten Jahren die Abfragesprache SPARQL[12] standardisiert, welche seit 2013 in der Version 1.1 vorliegt und von immer mehr Open Source- und kommerziellen Softwareanbietern implementiert und unterstützt wird. Der Umfang bestehender RDF-Datenbanken geht dabei immer weiter. Während vor einigen Jahren noch Datensätze von mehreren Milliarden Tripeln die obere Grenze bildeten, skaliert man heute bereits in den Bereich von Trillionen.[13]

[11] Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak.

[12] http://presentations.zazuko.com/SPARQL/

[13] Quelle : Präsentation Oracle an der ISWC 2015, Wien. https://twitter.com/FrankVanHarmele/status/654034761783083008

3 Studie

3.1 Linked Data Technologie

In diesem Kapitel wird genauer aufgezeigt, wie aus bestehenden Datenquellen RDF und Linked Data erzeugt werden kann. Die Prinzipien sind grundsätzlich für die meisten strukturierten Daten sehr ähnlich. Im Wesentlichen geht es dabei darum die Semantik der Informationen maschinenlesbar zu beschreiben und diese unter einem stabilen Identifikator (Link oder URI) bereit zu stellen. Dabei bedient man sich dem Technologie Stack des Web (HTTP) und entsprechenden Standards vom W3C rund um Semantic Web[14].

3.1.1 Technische Prinzipien für die Publikation und Nutzung von Linked Data

3.1.1.1 Publikation Linked Data

Die Daten werden oftmals dezentral in unterschiedlichsten Systemen gepflegt. Die Daten sind direkt (z.B. ODBC/APIs) oder indirekt (über NetCDF/CSV/XML/JSON Exports) verfügbar und werden durch einen spezifischen Transformator in RDF überführt. Die Transformation wird in Zusammenarbeit mit dem Datenhalter gemacht, dieser wird dabei durch RDF Spezialisten unterstützt. Der Prozess muss vollständig automatisierbar sein und bedarf einer entsprechenden Investition des Datenhalters.

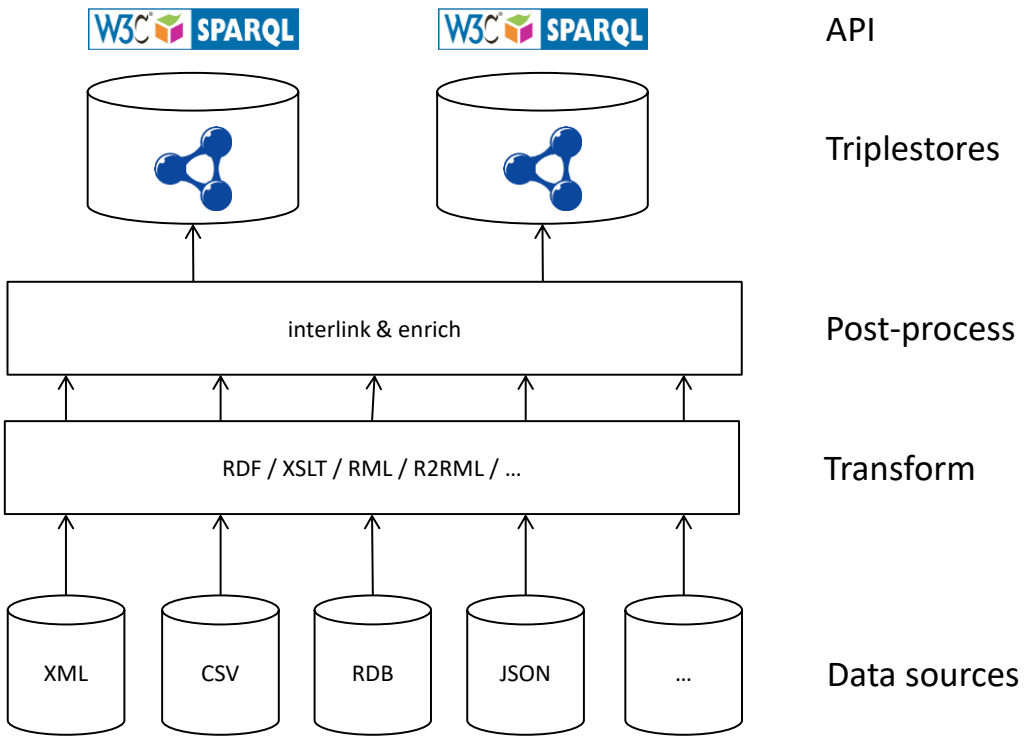


Abbildung 4: Process Linked Data

[14] https://www.w3.org/standards/semanticweb/

Nach der Überführung in RDF werden sie in einem zweiten Schritt mit anderen RDF Datensätzen verknüpft (interlinked & enriched).

Interlinking vereinfacht das Auffinden von weiteren, direkt in Verbindung stehenden Datensätzen. Dies geschieht in erster Linie über eindeutig referenzierbare Identifikatoren, die in den verschiedenen Datensätzen referenziert werden. Für den Datenkonsumenten soll dies das «follow your nose» Prinzip vereinfachen, womit den Links gefolgt wird und somit Zusatzinformationen aus weiteren Datensätzen einfach eingebunden werden können.

«Enriching» ist mit dem Interlinking verwandt, baut aber oft auf Analyse von Texten auf und versucht zum Beispiel durch Wort- und Satzanalyse (NLP) herauszufinden, welche Personen, Orte oder Stichworte in einem Text vorkommen. Diese Information kann wiederum mit den Originaldaten in Bezug gesetzt werden, was die Auffindbarkeit der Daten verbessern kann.

Im letzten Schritt, nach einer vorgängigen Qualitätskontrolle[15], werden die fertig aufbereiteten Daten in ihren jeweiligen Namensräumen in der Datenbank (Triplestore) abgelegt.

Die fertig aufbereiteten Daten werden den Datenkonsumenten über zwei Möglichkeiten zur Verfügung gestellt: Per HTTP API über den SPARQL Endpunkt und als Dump der Rohdaten in einer RDF Serialisierung.

RDF kann in vielen verschiedenen so genannten Serialisierungen abgelegt werden. Dabei gibt es Serialisierungen die für Maschinen sehr einfach zu lesen und schreiben sind wie Beispielsweise N-Triples. Andere wie Turtle eignen sich mehr für Menschen und sind entsprechend lesbar. In den letzten Jahren hat sich zudem JSON-LD etabliert, welches ein JSON-basierte Serialisierung von RDF ist und entsprechend oft im Web-Stack und von Webentwicklern verwendet wird. Jede Serialisierung kann dabei ohne Verlust in eine andere Serialisierung überführt werden.

3.1.1.2 Voraussetzungen

Eines der zentralen Ziele von RDF ist die Wiederverwendbarkeit der Daten. Um dies zu gewährleisten, sollten Daten relevante Metadaten entsprechend publiziert werden. Dabei gibt es folgende Voraussetzungen zu beachten:

- Identifikatoren sind klar definiert und nachvollziehbar (ein automatisierter Prozess generiert immer dieselben URIs aus den Identifikatoren für dasselbe Ding). Falls diese Identifikatoren nicht eindeutig sind, können Daten nur schwer oder gar nicht untereinander in Beziehung gesetzt werden.
- Namensraum definiert (möglichst nahe beim Datenhalter)
- RDF Schema/Ontologie definiert/formalisiert (bestehende Schemas wiederverwenden, neue wo nötig)

3.1.1.3 Überführen von MeteoSchweiz Daten nach RDF

In den Interviews bei MeteoSchweiz wurden in erster Linie zwei Datenquellen in den Vordergrund gestellt: Relationale Daten welche in Oracle gepflegt werden und hauptsächlich Punktbasiert sind und Vorhersagen in NetCDF welche in erster Linie Grid-basiert sind. Um relationale Daten nach RDF zu überführen eignet sich der Standard R2RML[16]. Dieser ermöglicht es, SPARQL Abfragen in

[15] „Interlinking“ und „Enriching“ stellen potenzielle Fehlerquellen für Linked Data dar, welche durch entsprechende Qualitätskontrollen zu minimieren sind.
[16] https://www.w3.org/TR/r2rml/

eine SQL Abfrage zu überführen. Damit können relationale Datenbanken über standardisierte Mappings nach RDF überführt werden. Diese Mappings können automatisch erstellt werden, um mit den Daten als Graph zu arbeiten. In der produktiven Nutzung empfiehlt sich allerdings ein manuell kuriertes Mapping, welches durch Domänenspezialisten erstellt und falls nötig gepflegt wird. Einige Datenbankhersteller bieten Unterstützung für R2RML, Oracle zum Beispiel im Produkt *Spatial and Graph*[17]. Zudem gibt es kommerzielle Produkte wie Capsenta Ultrawrap[18], welche mit entsprechender Optimierung der SQL Abfragen (durch Erstellen von Views) „near SQL speed" für SPARQL Abfragen erreichen. Um die Last auf dem relationalen System zu verringern kann auch automatisiert ein kompletter RDF Datendump erstellt werden, welcher danach in einem RDF Triplestore (RDF Datenbank) bereitgestellt wird und per SPARQL abgefragt werden kann. Es kommt dabei auf die konkreten Anforderungen an, welche Architektur schlussendlich sinnvoll ist.

NetCDF ist ein sehr effektives Format um grosse Datensätze mit Vorhersagen bereit zu stellen. Durch das binäre Datenformat lassen sich diese sehr stark komprimieren, gelesen werden sie mit entsprechenden Werkzeugen und APIs. NetCDF-4 basiert dabei auf HDF5, welches ebenfalls darauf spezialisiert ist grosse Datenmengen auszutauschen. Die beiden Datenformate sind Cube-orientiert und erlauben folglich das Speichern von Daten und Metadaten im selben Container. Ein automatisiertes überführen respektive generieren von RDF aus NetCDF-4 ist somit grundsätzlich möglich, auch wenn für die korrekte Beschreibung der Dimensionen und Messungen einige zusätzliche, Domänenspezifische Informationen nötig sind.

International gibt es einige Projekte, welche sich mit dem Überführen von NetCDF Daten nach RDF beschäftigen. Innerhalb dieser Studie wurde dabei CoverageJSON[19] und NetCDF-LD[20] genauer untersucht. NetCDF-LD wird unter anderem von CSIRO in Australien und UK Met Office vorangetrieben. Leider sind zum aktuellen Zeitpunkt noch keine brauchbaren Werkzeuge verfügbar, somit konnten die MeteoSchweiz Daten damit nicht automatisiert nach RDF überführt werden. Das vom *Reading e-Science Centre* des *Department of Meteorology* an der *University of Reading* propagierte CoverageJSON, überführt NetCDF Daten in ein JSON-Format. In der aktuellen, unfertigen Spezifikation ist vorgesehen die Daten in JSON-LD überführen zu können. Aktuell scheint dieser Teil aber noch nicht vollständig umgesetzt zu sein. Die Werkzeuge sind auch hier in einem sehr frühen Stadium und wurden deshalb ebenfalls nicht verwendet.

3.1.1.4 Consume

Die im Process-Schritt erstellten Daten werden per Tripelstore und SPARQL-Endpunkt zur Verfügung gestellt. Unterschiedliche Namespaces/Domänen lassen sich wie oben erwähnt auch zu jedem Zeitpunkt transparent in einen separaten Triplestore auslagern.

Der Consume-Layer teilt sich grundsätzlich in zwei Blöcke auf: Machine 2 Machine Kommunikation und ein Layer, der von Menschen konsumiert wird. Dabei baut der Human-Layer auf dem Machine 2 Machine-Layer auf.

[17] https://en.wikipedia.org/wiki/Oracle_Spatial_and_Graph
[18] https://capsenta.com/
[19] https://covjson.org/spec/
[20] https://medium.com/@jonyu8/netcdf-ld-and-why-its-going-to-be-awesome-321cea929bb5

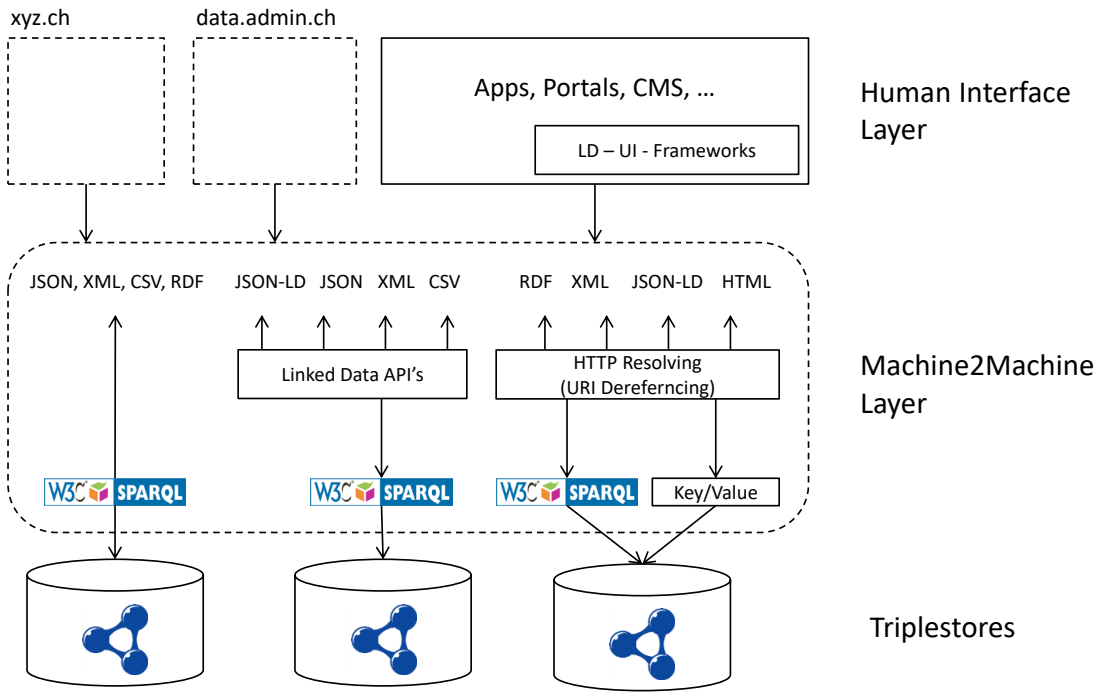


Abbildung 5: Consume Linked Data

3.1.1.5 SPARQL

Ein Linked Data-Service muss über SLAs dieselbe Verfügbarkeit erfüllen können, wie andere Kanäle auch. Dies um sicher zu stellen, dass sich Datenkonsumenten darauf verlassen können und darauf aufbauend produktive Services anbieten. Durch das Bereitstellen eines öffentlichen SPARQL-Endpunkts wird dem Datenkonsumenten ein sehr mächtiges und komplexes Werkzeug bereitgestellt. Dabei besteht das Risiko, dass der Datenkonsument Abfragen stellt, die sehr langsam sind oder die Datenbank im Extremfall temporär komplett blockieren (Denial of Service).

Aus diesem Grund ist es angebracht, mindestens zwei SPARQL-Endpunkte zur Verfügung zu stellen: Ein öffentlicher für explorative und komplexere Abfragen, die nicht über APIs bereitgestellt werden können/wollen/sollen und ein SPARQL-Endpunkt, der nur getestete Abfragen ausführt, welche zum Beispiel durch die API-Middleware gestellt werden. Im Kontext von MeteoSchweiz ist auch denkbar, dass zahlende Kunden ein besseres SLA erhalten als Kunden, welche auf Daten zugreifen die beispielsweise kostenlos sind.

Ein weiter Lösungsansatz ist dabei die so genannte horizontale Skalierung, wie sie im Kapitel 3.2 genauer erläutert wird.

3.1.1.6 HTTP Resolving / Dereferencing

Als Dereferencing bezeichnet man das Zugreifen per HTTP auf einen spezifischen Link (URI). Im Gegensatz zum Auflösen einer klassischen Webseite wird dabei nicht nur eine Antwort in HTML zur Verfügung gestellt, sondern die Maschine kann auch direkt die Rohdaten in einer RDF-Serialisierung verlangen.

Die HTML-Webseite erfüllt zwei Anforderungen: Dem Entwickler wird ein Werkzeug zur Verfügung gestellt, womit er die Daten einfach erforschen und interpretieren kann. Durch das Einbinden der

Daten in Rohform kann zudem einer Suchmaschine eine maschinenlesbare Form der Daten bereitgestellt werden. Dies lässt sich mit einem für Suchmaschinen verständlichen Format wie JSON-LD realisieren.

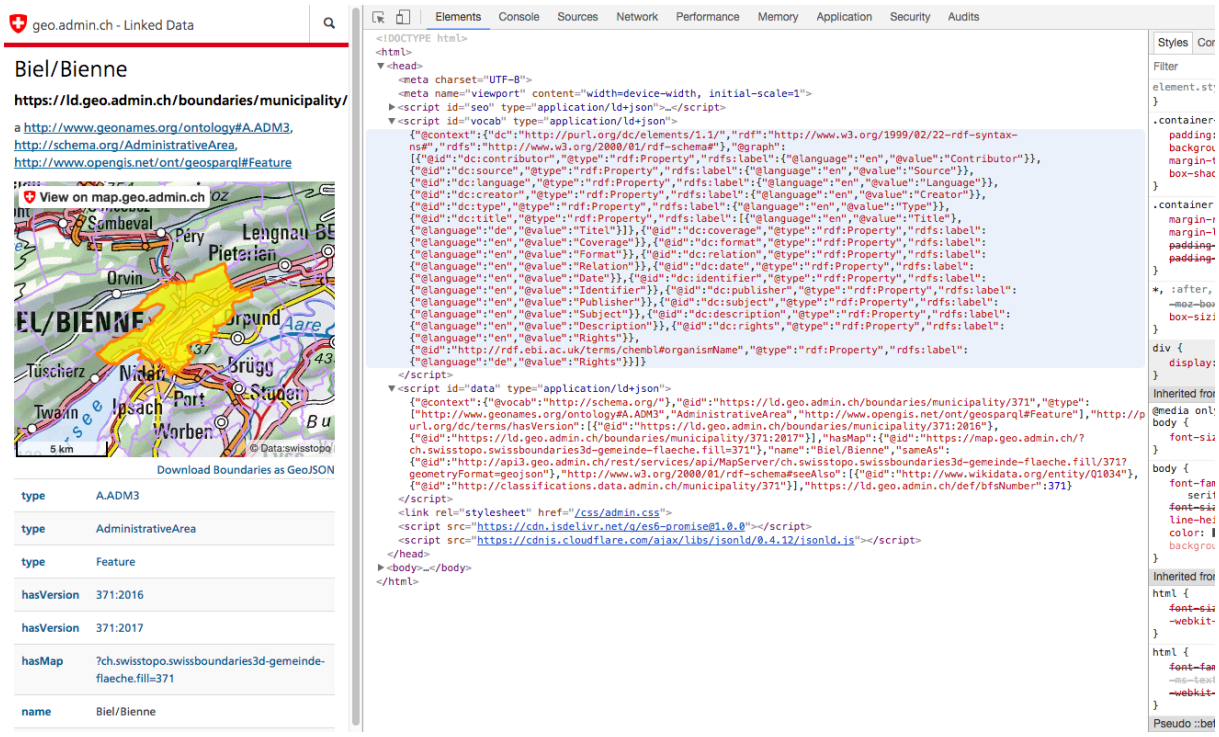


Abbildung 6: Eingebettete Metadaten in der Linked Data Version von Biel/Bienne bei Swisstopo

3.1.1.7 Linked Data APIs (API Middleware)

Linked Data APIs positionieren sich zwischen einfachem HTTP Dereferencing und SPARQL. Ziel ist es, die Hürde für nicht RDF-Spezialisten zu senken, damit die Daten in geeigneten Formaten konsumiert werden können.

Dabei gilt es eine Balance zu finden zwischen einfachen Schnittstellen (APIs) und einer Infrastruktur, die über lange Zeit unterhalten und garantiert werden kann. Da sich APIs mit den Daten weiter entwickeln werden und müssen stellt sich die Frage, wie diese entsprechend gepflegt werden können. Aktuell werden verschiedene Ansätze propagiert, verbreitet ist das Versionieren von APIs, wobei bei einer neuen Version unter Umständen sämtliche Anwendungen der API angepasst werden müssen. Dies führ bei Anwendern oftmals zu der so genannten API Hell[21], welche bestehende Applikationen unbrauchbar macht und/oder bei Entwicklern zusätzliche Unterhaltkosten erzeugt. Ein möglicher Lösungsansatz dafür sind so genannte Hypermedia-APIs[22]. Dabei beschreibt sich die Schnittstelle durch Metadaten selber, Clients lesen und interpretieren diese und passen sich entsprechend automatisch an. RDF und entsprechende Vokabulare bietet sich dabei als Beschreibungssprache an.

Ziel sollte sein, immer wieder gestellte Anfragen über solche APIs anbieten zu können, während dem Spezialfälle und Sonderwünsche zur Reduktion des Aufwandes nur per direktem SPARQL Zugriff bereitgestellt werden.

[21] https://t37.net/always-broken-inconsistent-and-non-versioned-welcome-to-api-hell.html

[22] Mike Amundsen - Twelve Patterns for Hypermedia Microservices. http://www.ustream.tv/recorded/102891495

Die API selber implementiert die Abfragen gegenüber dem Triplestore ebenfalls über SPARQL, HTTP Caching Proxies garantieren gegenüber dem Datenkonsumenten die nötige Zuverlässigkeit und Geschwindigkeit.

3.1.1.8 Human Layer

In diesem Layer soll die Information so aufbereitet werden, dass der Endanwender einen effektiven Nutzen aus den Daten ziehen kann („Kiosk“). Dies kann auf unterschiedlichen Stufen erledigt werden: Web- oder Mobileapplikationen, Portale oder CMS-Systeme, welche zum Beispiel Diagramme integrieren, welche aus den Daten und Abfragen erstellt werden.

Auf dieser Ebene wird nicht zuletzt der Datenhalter die aus seinem System erstellten Daten wieder konsumieren und in seine Systeme (z.B. öffentliche Webseite) integrieren. Dabei greift er wie jeder andere Datenkonsument auf die inzwischen stark verlinkten Daten zu und kann von den Synergien profitieren, die sich aus dem Interlinking & Enrichment Prozess ergeben haben.

Dies ist aktuell sicher der schwächste Punkt im ganzen Linked Data Technologiestack, da es nur wenig vielversprechende Ansätze gibt, die die Komplexität von Linked Data für Web- und App-Entwickler reduzieren. Eine entsprechende Unterstützung durch Frontend-Frameworks ist folglich angebracht.

In einem Prototyp mit dem Namen Uduvudu[23] wurden solche Ansätze innerhalb eines EU FP7 Forschungsprojektes erprobt. Uduvudu setzt dabei auf eine klare Trennung: Es gibt die Daten selber, eine sinnvolle Gruppierung derselben und ein entsprechendes Darstellen dieser Gruppierung. Daraus entstehen entsprechende Rollen: Datenhalter, Domänenspezialist und Interface-Spezialist. Jede dieser Rollen kann, aber muss nicht von denselben Personen übernommen werden. Somit kann sich jede Person auf ihre Fähigkeiten konzentrieren und entsprechend von den Fähigkeiten der anderen Rollen profitieren.

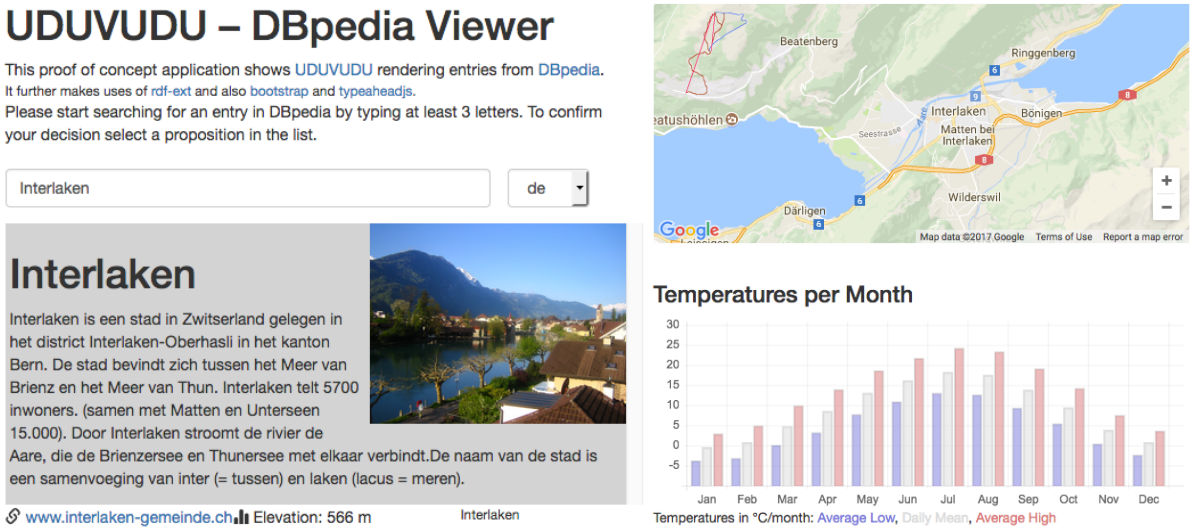


Abbildung 7: Automatische optimale Darstellung von Rohdaten zu Interlaken aus Wikipedia/DBpedia

Durch die in RDF enthaltene Semantik, kann die Maschine die Daten entsprechend interpretieren und in einem ersten Schritt die optimale Gruppierung finden. In einem zweiten Schritt werden die

[23] https://github.com/uduvudu/uduvudu

Daten möglichst optimal auf das Endgerät angepasst dargestellt. Die Agilität des RDF Datenmodells kann somit bis und mit Benutzerschnittstelle aufrechterhalten werden. Neue Daten können dargestellt werden, ohne dass sämtliche bestehenden Sichten manuell angepasst werden müssen.

3.1.2 Abgrenzung zu anderen Konzepten

Die nachfolgende Tabelle stellt wesentliche Aspekte anderer Datenhaltungskonzepte wie Relationale Datenbanksysteme (RDBMS), OLAP Würfel (Abfrageoptimierte logische Darstellung von Daten in mehreren Dimensionen) sowie statische Webseiten denen von RDF Stores (Linked Data) gegenüber.

Aspekt	RDF Store (Linked Data)	RDBMS	OLAP Würfel	Statische Webseiten
Datenstruktur	Netz / Graph in Triples	Verknüpfte Tabellen mit Datenintegrität	Stern / Schneeflocke	Hypertext, Links
Zugriff	URI	ODBC	ODBC, …	URL
Verknüpfung von Daten	intern / weltweit	intern	intern	intern / weltweit
Modellierung	Ontologie	ERD	ERD	Navigationskonzept
Aktualisierung	ETL / Query	Query, ETL	ETL	Upload
Abfrage	SPARQL, REST	SQL	SQL, …	HTTP
Visualisierung	SPARQL Endpoint, Applikation	Applikation	Analytics Tool	Browser

Tabelle 2: Abgrenzung zu anderen Konzepten

Ein RDF Store gehört zur Gruppe der «NoSQL» Datenbanken und dort zum Typ Graph-Datenbank (weitere Typen: Key-Value Store, Document Store, Column Store).

Weitere Klassifizierung: Persistierte Datenhaltung, in memory Datenhaltung.

3.2 Überlegungen zu Sizing / Betrieb

Meteo Daten werden in erster Linie als Dateien zur Verfügung gestellt. Die Dateien werden in sogenannten Datenpipelines erstellt und typischerweise auf ein System des Kunden hochgeladen. Diese Lösung hat zweifellos den Vorteil, dass sie relativ wenige Abhängigkeiten gegen aussen hat und zumindest in Bezug auf ähnliche oder gleiche Daten relativ einfach skaliert werden kann.

Ein RDF basierter Kiosk bedeutet daher initial einen technischen Mehraufwand.

Um Daten als RDF publizieren zu können, müssen zwei elementare Fragen geklärt sein:

- Was ist der Identifikator der Daten und

- was ist die Semantik der Daten.

Diese zwei Punkte müssen geklärt werden, um die Erstellung von RDF automatisieren zu können. Im Kontext von Meteodaten scheint in erster Linie der zweite Punkt noch Fragen aufzuwerfen. Es gibt aktuell keinen etablierten Standard, mit dem sich Meteodaten als RDF bereitstellen lassen. Alle bestehenden Projekte sind in einem relativ frühen Stadium. Im Proof of Concept wurden als Beispiel aus einer COSMO-1 Vorhersage auf einem kleinen Gebiet (Zürcher S-Bahn Gebiet) bis +33h der 2m-Temperatur in K (10-min. Auflösung) rund 2 Millionen Triples erzeugt. Diese werden alle 3 Stunden neu geliefert. Wenn die Daten für einen Tag abgelegt werden generiert dies alleine für diesen Datensatz rund 20 Millionen Triples.

Im Gegensatz zu relationalen Systemen hat RDF eine deutlich kürzere Geschichte vorzuweisen. SPARQL 1.1. ist erst seit 2013 ein Standard, die ersten Arbeiten begannen mit SPARQL 1.0 vor rund 10 Jahren. Trotzdem wurden in den vergangenen Jahren massive Fortschritte gemacht in Bezug auf Skalierung von grossen RDF basierten Datensätzen. Aktuelle Projekte wie UniProt stellen ihren Nutzern einen Öffentlichen SPARQL Endpunkt zur Verfügung mit aktuell 27 Milliarden Triples. Dasselbe Datenset wird mit zusätzlichen Daten auf einem Endpunkt mit rund 50 Milliarden Triples zur Verfügung gestellt.

Abbildung 8: UniProt Datensatz Statistik 2017, Quelle: https://twitter.com/uniprot/status/864770108455092224

Bisher wurde in Bezug auf Skalierung von RDF in erster Linie von vertikaler Skalierung gesprochen. Dabei wird dem Knoten/Rechner mehr Ressourcen zur Verfügung gestellt, zum Beispiel in Form von mehr Kernen oder mehr Memory. Bei RDF spielt dabei in erster Linie der Arbeitsspeicher eine grosse Rolle. Während man vor einigen Jahren noch einen Grossrechner von Cray brauchte um 2 Terrabyte Memory ansprechen zu können, werden solche Systeme heute von Firmen wie Dell zu bezahlbaren Preisen angeboten. Es ist davon auszugehen, dass diese Entwicklung auf absehbare Zeit nicht stoppen wird.

The following table shows recommended JVM memory and system memory requirements for Stardog.[15]

4. Table of Memory Usage for Capacity Planning

# of Triples	JVM Memory	Off-heap memory
100 million	3GB	3GB
1 billion	4GB	8GB
10 billion	8GB	64GB
20 billion	16GB	128GB
50 billion	16GB	256GB

Abbildung 9: JVM & System Memory Empfehlungen vom Stardog Triplestore. Quelle: http://www.stardog.com/docs/#_memory_usage

Grosse Datensätze wie Wikidata oder DBpedia zeigen zudem, dass SPARQL auch für viele Abfragen gleichzeitig skaliert werden kann. Nach aktuellen Statistiken hat Wikidata aktuell (2017) über 7 Millionen Abfragen pro Tag auf dem SPARQL Endpunkt, bei DBpedia sind es 6.2 Millionen pro Tag[24].

Aktuell arbeiten die ersten Hersteller an horizontaler Skalierung von SPARQL. Aus Sicht der Hardware sind dabei keine Grenzen gesetzt, man steigert die Leistung eines Systems durch Hinzufügen zusätzlicher Knoten. Stardog implementiert dazu SPARQL auf Basis von RocksDB, der Datenbank die den Social Graph von Facebook stemmt. Die Grenzen werden dabei in erster Linie durch die Parallelisierbarkeit der Software gesetzt. Stardog rechnet damit, dass sie damit bis Mitte 2018 hunderte von Milliarden Triples per SPARQL bereitstellen können. Die Pharmaunternehmen Merck experimentiert unter dem Projektnamen Halyard[25] mit einem ähnlichen Ansatz. Dieser Triplestore ist seit kurzem als Open Source Software verfügbar, wie gut sich damit arbeiten lässt werden die kommenden Monate zeigen. Im November 2017 hat Tomson Reuters ihren eigenen Knowledge Graph angekündigt[26], der intern vollständig auf RDF aufbaut. Der dafür eigens entwickelte Triplestore CM-Well ist ebenfalls als Open Source Software verfügbar und baut im Hintergrund auf einer hoch skalierbaren Plattform auf Basis von Apache Cassandra und Elastic Search auf. Laut Aussagen der Entwickler besteht dieser Graph aktuell aus 100 Milliarden aktiv verfügbaren Triples. Dazu kommt noch eine vollständige Versionierung der Daten, welche ebenfalls im Triplestore implementiert ist. Ende November 2017 hat Amazon die RDF Graph Datenbank Amazon Neptune vorgestellt[27], welche einen laut Ankündigung hoch skalierbaren RDF SPARQL Endpunkt in AWS bereitstellt. Einen etwas anderen Ansatz zur Skalierung wählt die Firma data.world[28]. Diese stellt einen hoch skalierbaren & hoch verfügbaren SPARQL Endpunkt bereit, der sich in erster Linie durch Zuverlässigkeit und günstige Preise von klassischen Triplestores unterscheidet. Im Hintergrund wird die seit länger als Open Source Software entwickelte Technologie und Serialisierung von RDF mit dem Namen

[24] https://twitter.com/linkedktk/status/869818744889110529
[25] https://merck.github.io/Halyard/
[26] https://www.thomsonreuters.com/en/press-releases/2017/october/thomson-reuters-launches-first-of-its-kind-knowledge-graph-feed.html
[27] https://aws.amazon.com/blogs/aws/amazon-neptune-a-fully-managed-graph-database-service/
[28] https://data.world/

HDT (Header dictionary triples) Verwendet. Dieses binäre RDF Format ist extrem kompakt, darauf aufbauend kann auch SPARQL bereitgestellt werden. Ein solcher SPARQL Endpunkt kann Datenmengen von 50 Milliarden Triples auf sehr günstiger Hardware mit nur wenig Speicher und günstigen CPUs bereitstellen. Entscheidend sind ausschliesslich schnelle SSD Harddisks, die ebenfalls für wenig Geld zu kaufen sind. Einziger Nachteil ist die Geschwindigkeit von so einem Endpunkt. Da nicht alle Daten in-memory sind, werden Abfragen in den meisten Fällen deutlich langsamer sein. Je nach Anwendungsfall muss das aber nicht ein grosser Nachteil sein und kann durch die günstigen Preise wettgemacht werden.

Zusammenfassend lässt sich festhalten, dass im Jahr 2017 RDF & SPARQL einen sehr grossen Schritt nach vorne gemacht hat. Nebst innovativen Startups kommen zunehmend grosse Player in den Markt, welche professionelle und hoch skalierbare Triplestores bereitstellen. Mit dem Eintritt von Amazon dürften auch die Preise sinken, um sehr grosse Datenmengen als RDF via SPARQL bereitzustellen. Zudem beginnen grosse Firmen sowohl intern wie in öffentlichen Datensätzen RDF und SPARQL zu verwenden. Viele dieser Firmen würden auch sehr gerne auf Meteodaten in RDF zugreifen können.

Ein weiterer Faktor der Skalierung ist Caching. RDF und SPARQL bauen beide auf HTTP auf, welches seit Jahren erfolgreich durch Caching-Server wie Varnish und Load Balancers wie HA-Proxy skaliert wird. So lange die Anfrage nicht ändert, kann folglich auch jedes SPARQL Query oder jeden direkten Zugriff auf eine RDF HTTP Ressource gecached werden. Für den Fall das übliche Abfragen statt direkt über SPARQL noch über eine einfachere HTTP API abstrahiert werden, gilt dies natürlich ebenfalls.

Es stellst sich grundsätzlich die Frage, ob alle Daten immer live zur Verfügung gestellt werden sollen. Gerade die hoch aufgelösten COSMO-1 Daten haben vermutlich eine relativ kurze Zeitspanne, in welcher sie für den Kunden relevant sind. Es wäre folglich gut vorstellbar, dass diese Daten nur während einer bestimmten Zeit im SPARQL Endpunkt live zur Verfügung gestellt werden. Falls jemand zu einem späteren Zeitpunkt «alte» Vorhersagen in SPARQL abfragen will, könnten diese beantragt und nochmals auf dem SPARQL Endpunkt bereitgestellt werden. Das System könnte diese beispielsweise als komprimierte (N-Triples) oder binäre (HDT) RDF Daten auf einem einfachen Dateisystem archivieren und auf Anfrage laden, wie das vorher erwähnte Beispiel data.world zeigt.

Um SLAs sicher zu stellen bietet es sich an, ein Cluster mit mehreren aktiven Knoten zu betreiben. Als Beispiel dazu dient der SPARQL Cluster LINDAS, welcher vom Schweizerischen Bundesarchiv BAR betrieben wird. Wird eine Lösung gewählt, die auf horizontaler Skalierung aufbaut, wird typischerweise mit bestehenden Cloud-Providern gearbeitet. So lässt sich die Leistung je nach Last aus- oder abbauen, was operative Kosten entsprechend optimiert.

3.3 Veränderung der Kundenbeziehung- und Bedürfnisse

Open Data zeigt, dass das Wiederverwenden von Daten in neuen Kontexten Mehrwert erzeugen kann. Eine Optimierung von bestehenden Geschäftsprozessen ist in Ländern wie der Schweiz elementar, um sich global weiterhin abgrenzen zu können. Wer unterschiedlichste Daten aus offenen wie internen Quellen effizient in eine neue Beziehung setzen kann, kann sich durch Mehrwissen von der Konkurrenz absetzen. Dies bedingt allerdings, dass die Daten möglichst ohne grossen Mehraufwand und maschinenlesbar in bestehende Systeme integriert werden können.

Eine der zentralen Fragen ist, wie Nutzer von Meteodaten von RDF basierten Technologien und Standards profitieren können. In Gesprächen mit Fachpersonen wurden zur Klärung dieser Frage die Bedürfnisse von bestehenden und möglichen zukünftigen Nutzern besprochen.

Kunden werden oftmals nur entsprechende Dateien geliefert, sei dies Messungen in CSV oder Vorhersagen in ASCII oder NetCDF. Selbst wenn einige der Kunden Datenlieferungen parametrisieren können, werden am Ende Dateien geliefert, eine direkte API für Programmierer gibt es folglich nicht.

Kunden sind oftmals nicht Domänenspezialisten. Mit Vorhersagen in NetCDF können sie wahrscheinlich eher wenig anfangen, wenn sie zudem Messdaten als CSV geliefert kriegen werden sie überfordert sein, weil im Gegensatz zur Vorhersage nur sehr wenige Messpunkte existieren.

Diese Probleme sind für spezialisierte Kunden sicher deutlich weniger relevant. NetCDF ist und bleibt ein wichtiges Format, um hoch aufgelöste Meteodaten auszutauschen.

RDF und Linked Data kann aber beiden Kundensegmenten – den Fachspezialisten als auch Domänenfremden - helfen, schnell und unabhängiger auf die für sie relevanten Daten zuzugreifen. Es wäre sicherlich sinnvoll, sämtliche Metadaten entsprechend aufzubereiten und als RDF bereit zu stellen. Mit entsprechend aufbereiteten Daten würde für Kunden einfach sichtbar, für welche zeitlichen und räumlichen Dimensionen welche Daten zur Verfügung stehen. Bei Messungen könnten zudem die vielen gepflegten aber oftmals für externe kaum zugänglichen Metainformationen zu Messstation bereitgestellt werden. Dies hilft bei der Interpretation der Daten und könnte auch etwaige Lücken oder Unstimmigkeiten erklären.

Einer der grossen Vorteile von RDF ist, dass die Problemstellung beim Erstellen von Daten nicht von Anfang an vollständig verstanden sein muss. Bei relationalen Systemen ist es deutlich schwieriger und aufwändiger, ein Schema zu einem späteren Zeitpunkt anzupassen[29]. Dies führt zum Beispiel dazu, dass Messdaten, welche aus technischen Gründen errechnet werden mussten, nicht als solche ausgewiesen werden können. Damit verliert man wertvolle Informationen, welche in spezifischen Fällen durchaus relevant sein können.

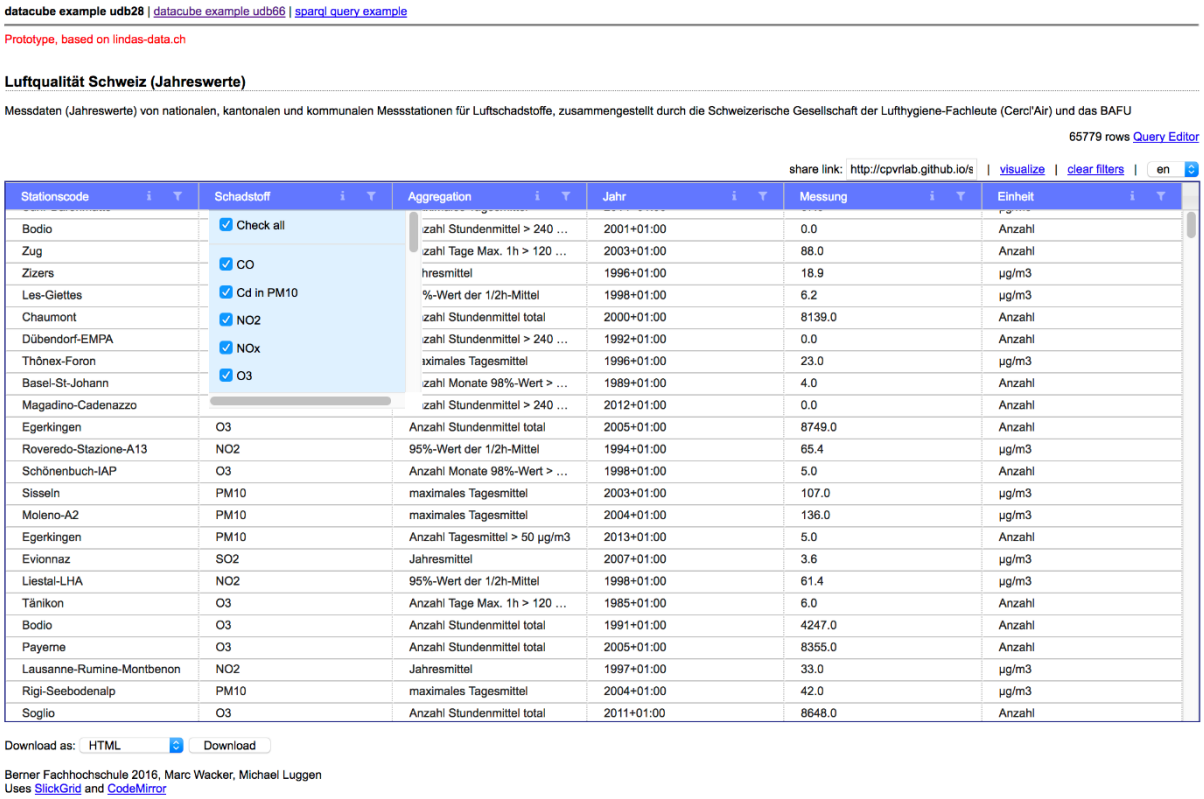
Mit RDF können die Daten und die entsprechenden Metadaten mit dem Problem wachsen, man spricht deshalb gerne von einem «Agilen Datenmodell». Bei einer Messstation könnten so zum Beispiel alle relevanten Informationen zum entsprechenden Sensor ebenfalls maschinenlesbar abgelegt werden. Damit ist es zu jedem späteren Zeitpunkt möglich zu erklären, wieso eine entsprechende Messung ab einem bestimmten Zeitpunkt feiner geworden ist. Solche Interpretationen sind heute nur möglich, wenn dem jeweiligen Kunden diese Informationen inkl. Erklärungen dazu bereitgestellt werden können. Dies ist sowohl für den Kunden wie für den Datenlieferanten ein entsprechender Mehraufwand. Natürlich kann mit RDF eine solche Interpretation nicht komplett der Maschine überlassen werden, aber es erleichtert zumindest das Auffinden solcher zusätzlichen Informationen und macht diese einfacher berücksichtigbar. Als konkretes Beispiel ist es heute nicht möglich direkt eine Abfrage zu machen, die nur Messdaten von Wetterstationen zwischen 1500 und 1800 Meter über Meer berücksichtig. Mit RDF können diese Informationen zu der Station abgelegt und folglich entsprechend abgefragt werden.

Nebst der Möglichkeit beliebige Metadaten gleich mit den Daten abzulegen, sticht in Bezug auf Meterologiedaten ein weiterer Punkt hervor: Geographische Informationen lassen sich mit den Vorhersagen oder Messungen ablegen und vor allem abfragen. Mit dem OGC (Open Geospatial Consortium) Standard GeoSPARQL[30] können diese Geoinformationen entsprechend direkt eingebunden werden. Dies kann den Zugriff auf Geobezogene Daten merklich vereinfachen, zumal typische Abfragen als Beispiele bereitgestellt werden können.

[29] https://www.societybyte.swiss/2017/02/22/datenkonsolidierung-wenn-das-relationale-modell-an-seine-grenzen-kommt/

[30] http://www.opengeospatial.org/standards/geosparql

Die in NetCDF vorliegenden Daten sind sehr Cube-orientiert[31] abgelegt. In RDF steht mit dem RDF Data Cube Vokabular ein Standard zur Verfügung, welcher sich immer mehr durchsetzt und unter anderem für Messreihen, Statistische Informationen und weitere zwei- bis mehrdimensionale Tabellen verwendet wird. Im erarbeiteten Use Case werden die COSMO Vorhersagen und die Sensordaten exemplarisch ebenfalls als RDF Data Cube modelliert.



Stationscode	Schadstoff	Aggregation	Jahr	Messung	Einheit
Bodio		zahl Stundenmittel > 240 …	2001+01:00	0.0	Anzahl
Zug		zahl Tage Max. 1h > 120 …	2003+01:00	88.0	Anzahl
Zizers		hresmittel	1996+01:00	18.9	µg/m3
Les-Giettes		%-Wert der 1/2h-Mittel	1998+01:00	6.2	µg/m3
Chaumont		zahl Stundenmittel total	2000+01:00	8139.0	Anzahl
Dübendorf-EMPA		zahl Stundenmittel > 240 …	1992+01:00	0.0	Anzahl
Thônex-Foron		ximales Tagesmittel	1996+01:00	23.0	µg/m3
Basel-St-Johann		zahl Monate 98%-Wert > …	1989+01:00	4.0	Anzahl
Magadino-Cadenazzo		zahl Stundenmittel > 240 …	2012+01:00	0.0	Anzahl
Egerkingen	O3	Anzahl Stundenmittel total	2005+01:00	8749.0	Anzahl
Roveredo-Stazione-A13	NO2	95%-Wert der 1/2h-Mittel	1994+01:00	65.4	µg/m3
Schönenbuch-IAP	O3	Anzahl Monate 98%-Wert > …	1998+01:00	5.0	Anzahl
Sisseln	PM10	maximales Tagesmittel	2003+01:00	107.0	µg/m3
Moleno-A2	PM10	maximales Tagesmittel	2004+01:00	136.0	µg/m3
Egerkingen	PM10	Anzahl Tagesmittel > 50 µg/m3	2013+01:00	5.0	Anzahl
Evionnaz	SO2	Jahresmittel	2007+01:00	3.6	µg/m3
Liestal-LHA	NO2	95%-Wert der 1/2h-Mittel	1998+01:00	61.4	µg/m3
Tänikon	O3	Anzahl Tage Max. 1h > 120 …	1985+01:00	6.0	Anzahl
Bodio	O3	Anzahl Stundenmittel total	1991+01:00	4247.0	Anzahl
Payerne	O3	Anzahl Stundenmittel total	2005+01:00	8355.0	Anzahl
Lausanne-Rumine-Montbenon	NO2	Jahresmittel	1997+01:00	33.0	µg/m3
Rigi-Seebodenalp	PM10	maximales Tagesmittel	2004+01:00	42.0	µg/m3
Soglio	O3	Anzahl Stundenmittel total	2011+01:00	8648.0	Anzahl

Abbildung 10: Beispiel eines generischen RDF Data Cube Browsers, angewendet auf BAFU Umwelt Basisdaten

Australien stellt mit dem ACORN-SAT Datenset[32] die täglichen Temperaturen über 100 Jahre als Linked Data zur Verfügung. Beim Bund sind verschiedene Behörden daran, RDF Data Cubes zu erstellen oder dies in Betracht zu ziehen. Prototypische Umsetzungen gibt es bisher vom Bundesamt für Statistik, vom Bundesamt für Umwelt (BAFU) und von Statistik Stadt Zürich (SSZ). International nimmt der Trend für RDF Data Cubes ebenfalls stark zu, beispielsweise bei Statistics Scotland[33] und OpenBudgets[34].

Daten als RDF Data Cube zur Verfügung zu stellen erleichtert zudem die Integration dieser Daten in analytische Werkzeuge, wie zum Beispiel RAWGraphs[35] oder Tableau[36](siehe Abbildung 11 - Abbildung 13). Im Gegensatz zu pur CSV basierten Werkzeugen kann der Domänenspezialist durch die Paging-Funktionalität von SPARQL ohne Performance-Probleme durch grosse Zeitreihen oder

[31] https://de.wikipedia.org/wiki/OLAP-W%C3%BCrfel

[32] http://www.semantic-web-journal.net/system/files/swj1313.pdf

[33] http://statistics.gov.scot/

[34] http://openbudgets.eu/

[35] http://rawgraphs.io/

[36] https://www.tableau.com/

ähnliche Dimensionen navigieren. Dies vereinfacht den Zugang zu den Meteo-Rohdaten und macht diese somit einem neuen Publikum zugänglich.

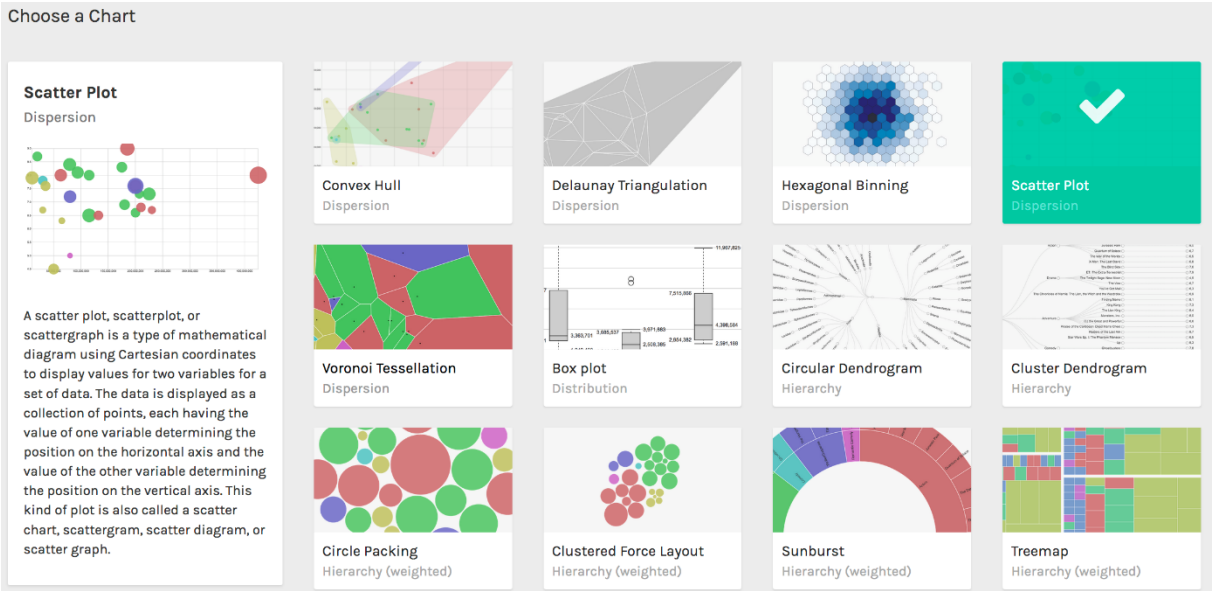


Abbildung 11: RAWGraphs Auswahl der möglichen Visualisierungen

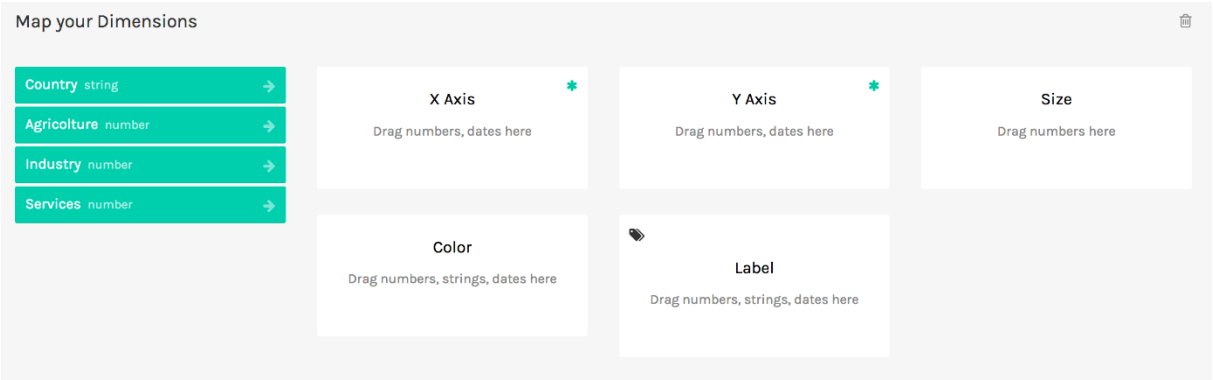


Abbildung 12: Auswahl der Dimensionen der Daten in RAWGraphs

Im Gegensatz zu NetCDF und CSV basierten Datenzugriffen ist ein Zugang über SPARQL immer ein Zugriff auf einen Webservice. Dies erhöht zwar die technischen Anforderungen an eine Plattform aber lässt auch eine deutlich aussagekräftigere Statistik zu über welche Daten wie stark verwendet werden. Bei offline-Formaten wie NetCDF oder CSV kann nach dem Download nicht mehr nachvollzogen werden, wie oft und in welcher Form die Daten verwendet werden.

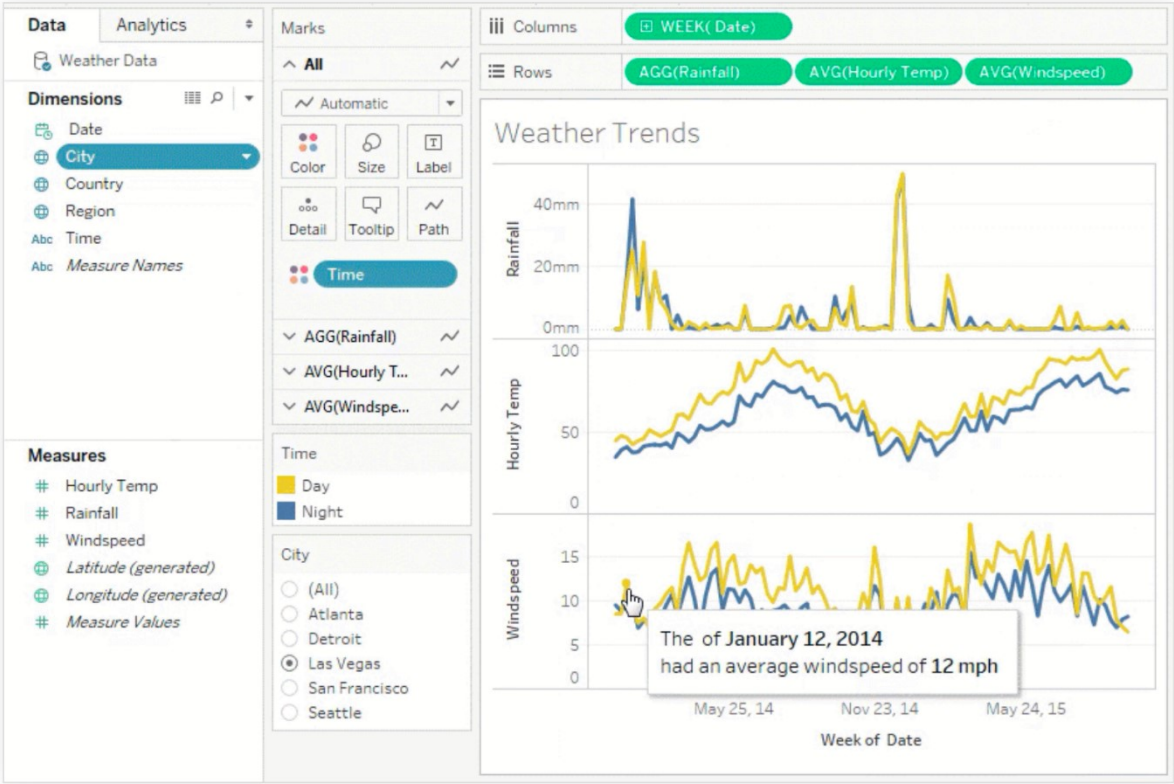


Abbildung 13: Wetterdaten ausgewählt auf Basis einer Auswahl in Tableau (Quelle: https://www.tableau.com/products/desktop)

Ein weiterer Vorteil von RDF und Linked Data ist deren dezentrales Konzept. Im Gegensatz zu anderen Technologien geht RDF davon aus, dass ein grosser Teil der Daten nicht zwingend in derselben Datenbank gepflegt und/oder abgefragt werden kann oder muss. Datenhalter können folglich auf andere Referenzdatensätze verweisen, die vom jeweiligen Datenhalter gepflegt und als RDF bereitgestellt werden. Statt die Daten nochmals im eigenen Datensilo zu duplizieren, kann einfach auf die jeweiligen Links (URIs) der Daten verwiesen werden, wie dies bei einer Webseite geschieht. Dies vereinfacht das Integrieren von anderen Referenzdatensätzen enorm und vereinfacht die Datenhaltung und Pflege, da diese immer vom Datenhalter übernommen wird.

In der Schweiz werden als Beispiel Daten vom Bundesamt für Statistik bereitgestellt, welche das historisierte Gemeindeverzeichnis in RDF abbilden. Swisstopo stellt dazu die Umrisse der Gemeinden zur Verfügung. Damit lässt sich eine kombinierte Abfrage erstellen, welche aus den beiden SPARQL-Endpunkten die relevanten Informationen abfragt. Mit entsprechenden Plugins können diese ohne irgendwelche Programmierung auf einer Karte visualisiert werden.[37]

[37] Dieses und weitere Beispiele können direkt ausgeführt werden: https://github.com/zazuko/fso-lod/blob/master/doc/eCH0071/sparql.md

```
16  SERVICE <http://data.admin.ch/query> {
17    ?event a gont:MunicipalityChangeEvent ;
18      gont:date ?date ;
19      gont:id ?eventid .
20
21    ?munabo a gont:MunicipalityVersion ;
22      gont:longName ?abolitionName ;
23      gont:municipality ?munuri ;
24      gont:abolitionMode ?abmode ;
25      gont:abolitionEvent ?event .
26      ?abmode skos:prefLabel ?abolitionmodeName .
27  }
28  FILTER ( ?date >= "2016-01-01"^^xsd:date && ?date <= "2016-12-31"^^xsd:date )
29
30  ?shapeuri a <http://www.geonames.org/ontology#A.ADM3> ;
31    dct:hasVersion ?version16 ;
32    rdfs:seeAlso ?munuri .
33
34  ?version16 geo:hasGeometry ?Geometry .
35  ?Geometry geo:asWKT ?Coords  .
```

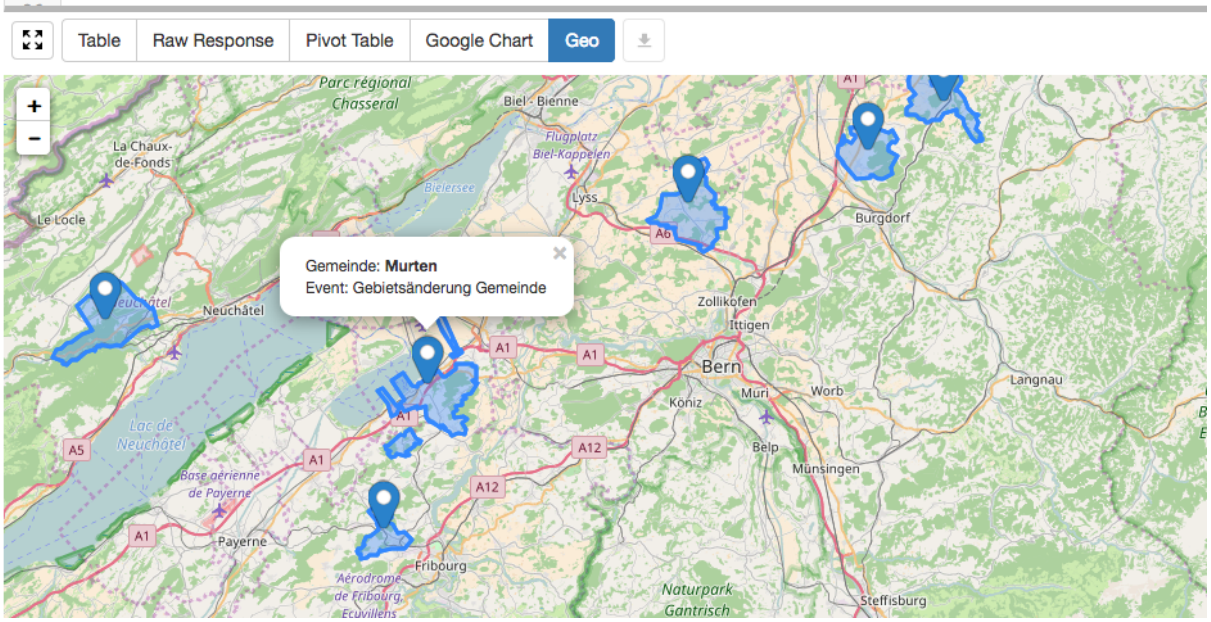


Abbildung 14: SPARQL Abfrage der aufgehobenen Gemeinde im Jahr 2016 mit entsprechender Visualisierung der Gemeinde. Die Umrisse der Gemeinde können direkt in der SPARQL-Konsole dargestellt werden, in diesem Beispiel aufbauend auf Openstreetmap Karten. Für Swisstopo wurde ein Plugin entwickelt, welches die Schweizer Karten in diese Sicht integriert.

In den geführten Interviews mit Fachspezialisten wurde deutlich, dass sich diese und ähnliche Probleme mit Datenduplizierung- und Pflege auch bei Meteodaten stellen. Solche Doppelspurigkeiten könnten mit RDF zumindest deutlich reduziert werden.

Auch die Führung von mehrsprachigen Informationen wird mit RDF stark vereinfacht, da das Datenmodell Mehrsprachigkeit von Grund auf beherrscht und bis und mit Abfragesprache SPARQL konsequent unterstützt wird. In der Schweiz ist dies bekanntlich ein nicht zu unterschätzender Vorteil der mit bestehenden Technologien oftmals nur schwer umsetzbar ist.

Eine Art Datenkiosk - im Sinne eines Selbstbedienungsladens - würde sich als Plattform für die Bereitstellung von Meteodaten sehr gut eignen. Der Technologiestack rund um RDF und Linked Data bietet ein ideales Fundament, um so einen Datenkiosk aufzubauen. SPARQL dient dabei als

Abfragesprache auf dem untersten Level, Abstraktionen und Vereinfachungen könnten darauf aufbauend APIs bieten welche häufig getätigte Abfragen für die Datenkonsumenten vereinfachen. Darauf aufbauend könnten Werkzeuge wie Tableau oder RAWGraph oder Visualisierungen auf Basis von D3[38] und d3.express[39] die technische Basis für einen neuen Zugriff auf Meteodaten liefern. Initial würde der Kiosk sinnvollerweise Kunden die Daten einfacher und besser zugänglich machen, mittelfristig könnte ein komplett neues Segment von Kunden angesprochen werden und somit die Verbreitung von Meteodaten erhöht werden.

3.4 Wissensbedarf & Maturitätslevels

In Kapitel 2.1 wird das 5 Sterne Modell von Tim Berners Lee ausführlich behandelt. Dieses zeigt die 5 Stufen der Maturität des Datenpublizierens. Jede Stufe stellt dabei eine messbare Verbesserung bezüglich maschinenlesbarer und verknüpfter Nutzung dar. Ab Stufe 4 wird technisch auf RDF gesetzt.

Ziel müsste es generell sein, die Stufen 4 und 5 anzupeilen. Dafür sind entsprechende Rollen und Skills sowohl Seitens des Bereitstellers von Daten als auch Seitens der Datennutzer notwendig. Die Nachfolgende Tabelle, beschreibt mögliche Rollen und deren Einsatz.

Rolle	**Beschreibung / Skills**	**Bereich**	**Bereitsteller-seitig oder Nutzerseitig**
Datennutzer ohne spezielles Fachwissen	Allgemeinheit. Zugriff auf Daten über Applikation, Kiosk, etc. Keine Linked Data spezifischen Skills	Datennutzung	Nutzerseitig
Datennutzer mit speziellem Fachwissen im jeweiligen Fachbereich	Luftfahrt, Wissenschaft, Behörden. Zugriff auf Daten über Applikation, Kiosk, e-Shop, vorproduzierte Datensätze als Dateidownload Fachwissen, keine Linked Data spezifischen Skills	Datennutzung	Nutzerseitig
Linked Data User (SPARQL)	Interpretiert die Ontologie und fragt die Daten mittels SPARQL-Queries ab für die Weiterverarbeitung in (Web-)Applikationen, Machine Learning, komplexen BI-Auswertungen, etc. Skills und Erfahrung im Bereich SPARQL notwendig.	Datennutzung	Bereitsteller-seitig

[38] https://d3js.org/

[39] https://medium.com/@mbostock/a-better-way-to-code-2b1d2876a3a0

Linked Data Designer	Gestaltet die Ontologie/Vokabular, die Datenmappings und den Import von Daten aus Umsystemen in den LD Store. Umfassende Skills und Erfahrung im Bereich RDF, SPARQL und der Konzepte von Linked Data im Allgemeinen notwendig.	Datenpublikation	Bereitstellerseitig
DB Administrator / DB Engineer	Kennt die bestehende Applikationslandschaft, Datenbanken. Unterstützt bei der Interpretation der Daten und dem Mapping der Strukturen in eine Ontologie. Kennt ETL Tools und kann Daten für die Serialisierung aufbereiten. DBA, DBE Fachwissen mit Interesse für Linked Data.	Datenpublikation	Bereitstellerseitig

Tabelle 3: Rollen

3.5 Erkenntnisse aus dem Proof of Concept (Show Case)

Im Rahmen der vorliegenden Studie wurde ein zeitlich begrenzter Show Case durch die Universität Bern umgesetzt mit dem Ziel, den Nutzen von Linked Data, als Variante für die Datenpublikation von Wetter- und Klimadaten zur einfachen Weiternutzung und Integration mit weiteren Datensätzen (in diesem Showcase Daten der SBB) aufzuzeigen.

Der Show Case besteht aus der Publikation von Datensätzen der MeteoSchweiz sowie der SBB plus einer Applikation, welche die publizierten Daten einerseits zur Steuerung der Heizung einer Zugkomposition der SBB nutzt und andererseits die dadurch eingesparte Energie visualisiert.

Weiter dient der Show Case als Proof of Concept für diverse Punkte aus der Studie.

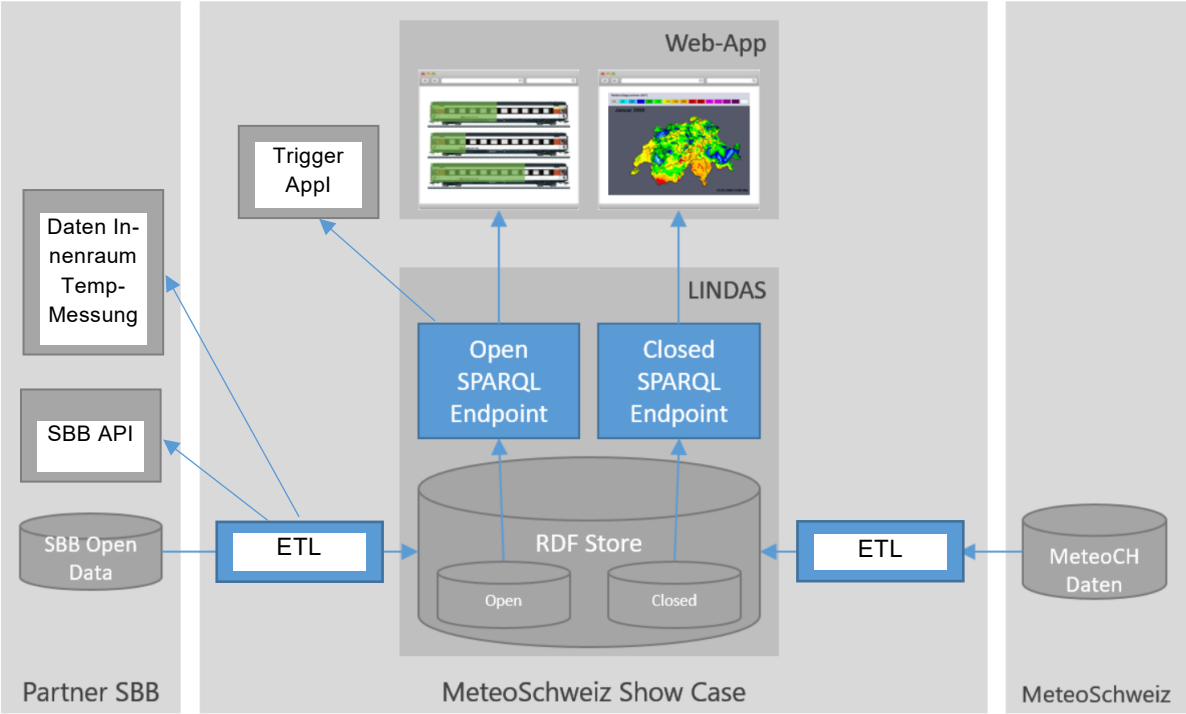


Abbildung 15: Konzeptionelle Sicht auf den Show Case

Die Daten von MeteoSchweiz und SBB werden laufen aufbereitet und importiert. Die Trigger App sowie die open und closed Web-Applikationen greifen für die Datenabfrage auf einen SPARQL-Endpoint zu.

Im Rahmen der Diskussionen zum Show Case sind noch viele weitere mögliche Anwendungsfälle aufgetaucht. Dies zeigt, dass das Interesse sehr gross und das Potential immens und vielfältig ist.

3.5.1 Look and Feel der Webapplikation

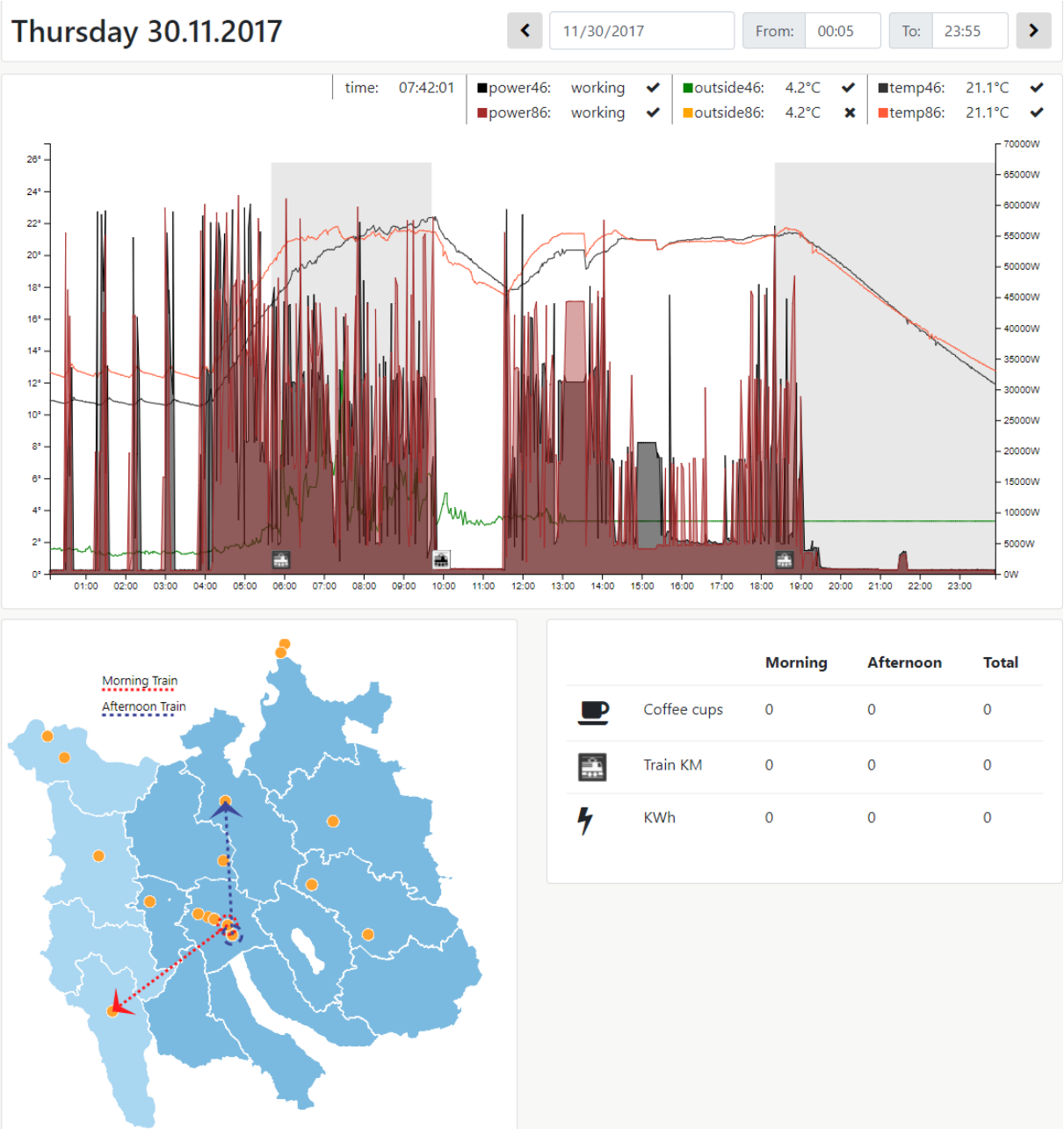


Abbildung 16: Tagesansicht

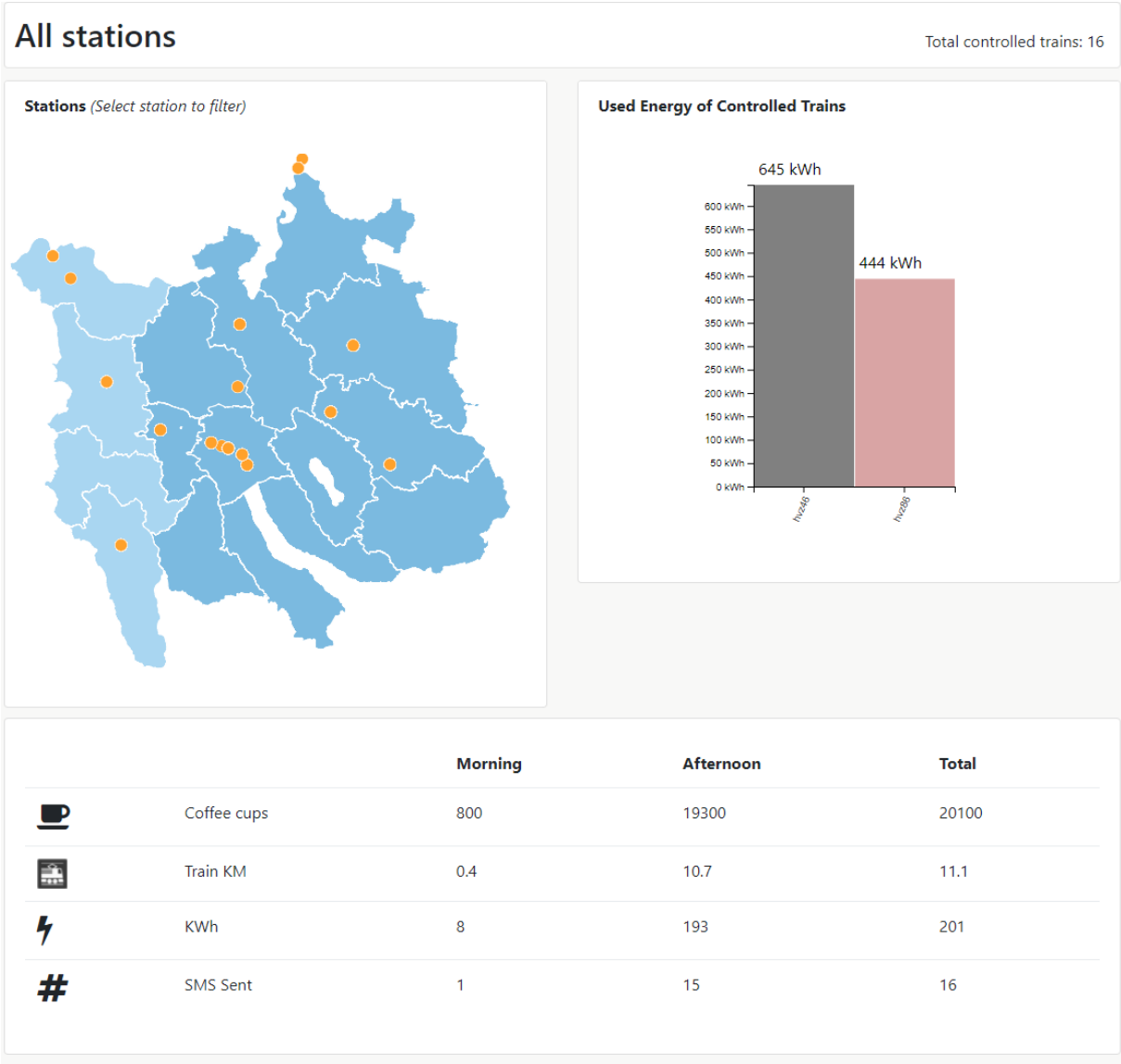


Abbildung 17: Zusammenfassung mit kumulierten Einsparungen

3.5.2 Authentifizierung open/closed Data

Eines der Ziele des Show Case war es, die Möglichkeiten aufzuzeigen wie mittels Linked Data sowohl «Open Data» als auch «Closed Data» zu zugänglich gemacht werden könnte.

Prinzip: Daten auf dem SPARQL End Point sollen entweder öffentlich oder nur mit Benutzername / Passwort zugänglich sein.

Der Zugang zu geschlossenen Daten (sogenannten Closed Graphs) wurde durch Authentifizierung auf dem Triple Store gelöst. Der Zugriff auf den Triple Store ist dabei unterschiedlich. Für den Zugang zu offenen Daten, wird direkt ein GET Request auf https://test.lindas-data.ch/sparql/ gesendet. Für Daten aus dem Closed Graphen wird die Stardog npm library (https://www.npmjs.com/package/stardog) verwendet.

Alle Daten, welche für die Steuerung gesammelt werden, werden in zwei unterschiedliche Graphen geschrieben. In den einen Graphen werden die Logs der Steuerungsapplikation und die Informationen, wann und ob die Heizung des Zuges angesteuert wurde geschrieben. In den zweiten Graphen

werden die Meteodaten (Temperatur und Globalstrahlung), die für die Berechnung der SMS Sendezeitpunkte verwendet wurden gespeichert. Im Gegensatz zu den kompletten Meteodaten, welche täglich von MeteoSchweiz auf dem Endpoint publiziert werden, wird die Kopie nicht nach 24 Stunden gelöscht. Diese Daten sind dann nur noch auf dem closed Endpoint verfügbar. Die Daten der SBB (Zugnummer, An-/ Abfahrtszeitpunkte/ -orte, Messung der Innentemperatur und Energieverbrauches des angesteuerten und des Kontrollwagens, Messung der Aussentemperatur (SBB)) sind alle open.

In der Webapplikation unterscheidet sich die verwendete Datengrundlage dann je nachdem ob der Benutzer eingeloggt ist oder nicht und somit sind mehr oder weniger Daten angezeigt.

4 Weiterführende Informationen

4.1 Vorhandene Vokabulare im Bereich Meteorologie

Im Bereich Meteorologie sind gemäss aktuellem Kenntnisstand keine spezifischen Vokabulare vorhanden. Im Verlauf dieses Projektes wurden verschiedene Erweiterungen in Bezug auf das RDF Data Cube Vokabular publiziert, welches auch in diesem Projekt verwendet wird. Die Publikationen sind allerdings erst nach der Umsetzung des ShowCase gemacht worden, aus diesem Grund wurden die Empfehlungen aus den beiden Erweiterungen noch nicht berücksichtigt.

Die beiden Erweiterungen sind *QB4ST: RDF Data Cube extensions for spatio-temporal components*[40] und *Publishing and Using Earth Observation Data with the RDF Data Cube and the Discrete Global Grid System*[41].

Die erste Erweiterung (QB4ST) ist vom *Open Geospatial Consortium (OGD)* und der *Spacial Data on the Web Working Group* und zeigt auf, wie Daten in der räumlich-zeitlichen Ebene im RDF Data Cube Vokabular modelliert werden können und wie dafür spezielle Referenzsysteme (Beispielsweise Koordinatensysteme) in RDF beschrieben werden können.

Die zweite Erweiterung bezieht sich ziemlich genau auf das, was in diesem Projekt umgesetzt wurde. Es beschreibt, wie dichte räumliche Rasterdaten als RDF Data Cube beschrieben werden können und wie Metadaten entsprechend aufbereitet werden können. Dieses Dokument ist ebenfalls vom OGD publiziert worden und nimmt Bezug auf QB4ST. Die Autoren selber kommen von der Australian National University und sind wahrscheinlich im Kontakt mit dem australischen Wetterdienst. Das Dokument fokussiert in erster Linie auf einer sehr ausführlichen Beschreibung der Metadaten. Dazu werden verschiedene andere Ontologien und Vokabulare wiederverwendet und gezeigt, wie sie in Bezug auf RDF Data Cubes modelliert werden können. Dies geht soweit das man in RDF beschreiben kann, welcher Sensor oder welche Maschine welchen Wert gemessen respektive errechnet hat.

Die Autoren gehen davon aus, dass es aktuell nicht möglich ist die eigentlichen Daten als RDF zur Verfügung zu stellen. Der Autor dieser Zeilen würde dem widersprechen, dieses Projekt zeigt, dass dies durchaus gemacht werden kann und eine Frage der entsprechenden Triplestores/Implementierung ist.

Wie erwähnt wurden die Empfehlungen aus diesen beiden Dokumenten im ShowCase noch nicht umgesetzt. Es zeigt aber klar, dass aktuell verschiedene Gruppen an ähnlichen Projekten mit ähnlichen Zielrichtungen arbeiten.

[40] https://w3c.github.io/sdw/qb4st/

[41] https://www.w3.org/TR/eo-qb/

4.2 Bekannte Vorhaben im Umfeld von Linked Data

4.2.1 Schweizer LD Projekte

Data provider	Dataset	Publication_status	URL_description_page
Federal Office for Civil Protection, PCP Section	PCP Inventory	Planned for December 2016	http://make.open-data.ch/wiki/data:glam_ch#pcp_inventory
Canton of Zurich	Inventory of Historical Monuments of the Canton of Zurich	Planned for early 2017	
City of Zurich	Inventory of Historical Monuments of the City of Zurich	Planned for early 2017	
aLOD Konzortium	Öffentliches Datenset vom Bundesarchiv und einigen Kantonen respective Städten.	Public	http://data.alod.ch/
OpenGLAM CH	Swiss GLAM Inventory	Public	http://make.open-data.ch/wiki/data:glam_ch#swiss_glam_inventory
Theatersammlung	Datensätze aus verschiedenen Theater- und Tanzarchiven	Öffentlich	http://make.open-data.ch/wiki/data:glam_ch#swiss_theatre_metadata
Plazi.org	Plazi Treatment Bank	Planned for end 2016	
Swissbib	Swiss Library Metadata	Publication planned for April 2017 (test version possibly earlier)	
sib.swiss			
LINDAS use-cases datasets	Verschiedene Datensätze aus dem LINDAS Pilot	Öffentlich	https://github.com/zazukoo/lindas-datasets/
Stadt Zürich	Statistik Stadt Zürich, Ersatz für das bestehende Jahresbuch. Ab 2018 nur noch digital und Linked Data getrieben.	Produktion ab Frühling 2018	

Bund	TERMDAT, ein umfangreiches mehrsprachiges Fachwörterbuch vom Bund.	Geplant für Ende 2017	https://www.termdat.bk.admin.ch/Search/Search
Bundesamt für Statistik	Historisiertes Gemeindeverzeichnis	Öffentlich	http://data.admin.ch/search/
Swisstopo	Linked Data Dienst: GeoDaten semantisch verlinken	Öffentlich	http://ld.geo.admin.ch/
Bundesamt für Verkehr/SBB	DIDOK, ÖV Betriebspunkte in der Schweiz	Öffentlich	http://lod.opentransportdata.swiss/
Bundesamt für Umwelt	BAFU UBD 28 & 66: Exemplarische Data Cubes von BAFU Messwerten	Öffentlich	https://github.com/lindas-uc/bafu_ubd

Tabelle 4: Linked Data Projekte Schweiz

4.2.2 Internationale Projekte

Im Folgenden eine Sammlung von Business Use Cases für Linked Government Data im internationalen Umfeld ohne Anspruch auf Vollständigkeit:

Titel: Informationsmanagement bei der EC - Publication Office & Council of European Union

Beteiligte Stakeholder: Council of E.U.; EC Publication Office

Kurzbeschreibung Use Case: Mittels der Verlinkung von diversen Metadatenbeständen im Publication Office (PO) mit Beständen des Councils (Rat der EU) werden Daten und Dokumente kontextualisiert und

damit für alle MitarbeiterInnen der Abteilungen der Europäischen Kommission und des Councils (Rat der EU) besser erschließbar gemacht. Teile werden auch mittels LOD Publishing beim Council publiziert.

Verwendete Daten: EuroVoc, MRD (MetaDataRepository der EC: http://publications.europa.eu/mdr/), Dokumente Publication Office, Dokumente Council of European Union.

Link, Kontakt: Internes System nicht verfügbar, LOD Publishing: http://data.consilium.europa.eu/

(a)**Titel:** Westtoer

Beteiligte Stakeholder: Tourismus Flandern, regionale Regierung, TenForce (Umsetzer)

Kurzbeschreibung Use Case: Touristische Informationen (der Tourismusbüros der Region) mittels linked data verknüpft und damit intern in Verwendung, zusätzlich publiziert als LOD (Data Hub). In Planung: Tourismus Vokabular für die Region als LOD (SKOS).

Verwendete Daten: Tourismusdaten, Geodaten, Tourismus-Vokabular

*(i)**Link, Kontakt:** Westtoer: westtoer.be, datahub.westtoer.be*

(b)Titel: Education Service Australia - School Online System / Thesaurus (ScoT)

Beteiligte Stakeholder: Education Service Australia (ESA)

Kurzbeschreibung Use Case: ESA (http://www.esa.edu.au/) hat den School Online Thesaurus (ScOT) als Linked Open Data erarbeitet und verwendet diesen zur Klassifikation sämtlicher Lerninhalte an Australiens Schulen. Dieser LOD Thesaurus wird also intern für die Verwaltung der Lehr- und Lern-Repositories verwendet und ist extern zugänglich, um Themen zu erschließen und alle zugehörigen relevanten Lehr- und Lerninhalte darüber zu erhalten.

Verwendete Daten: School Online Thesaurus, gesamte Lehr- und Lerninhalte Australiens.

*(i)**Link, Kontakt**: http://scot.curriculum.edu.au/ und http://vocabulary.curriculum.edu.au/*

*(ii)**Titel:** UK Government Organogramms*

Beteiligte Stakeholder: alle relevanten Abteilung der UK Regierung

Kurzbeschreibung Use Case: Darstellung der Organigramme des UK Governments inklusive Hierarchien und Bezügen. Applikation wird intern und extern verwendet.

Verwendete Daten: Organigramme aller Abteilungen des UK Governments und Verdienstmodelle mit konkreten Bezügen aller Abteilungen und Hierarchien.

*(iii)**Link, Kontakt**: http://data.gov.uk/organogram/cabinet-office*

(c)Titel: Amsterdam Fire Department

*(i)**Beteiligte Stakeholder**: Amsterdamer Feuerwehr*

Kurzbeschreibung Use Case: durch die Bereitstellung und Verlinkung von Geodaten und Echtzeitdaten aus der Telematik und Daten der Lage und Beschaffenheit der Grachten wurde es der Amsterdamer

Feuerwehr ermöglicht, Einsätze rascher und präziser durchzuführen, da Anfahrt und Vorortarbeiten dadurch unterstützt werden.

Verwendete Daten: Pager Daten, Geodaten, Einsatzdaten, Telematikdaten.

(d)Link, Kontakt:
http://www.dataversity.net/keynote-video-and-updates-from-the-amsterdam-fire-department/
http://blog.kasabi.com/2012/01/31/real-time-data-from-amsterdam/
http://de.slideshare.net/semanticfire/iswc2014-devworkshopbartvanleeuwen
http://www.epsiplatform.eu/content/amsterdam-fire-brigade-linked-data

(e)Titel: openlaws.eu

Beteiligte Stakeholder: openlaws GmbH & FP7 Projekt gefördert von DG JUSTICE (EC), Kooperation mit EuroLex, RIS (Rechtsinformationssystem Österreich).

Kurzbeschreibung Use Case: openlaws.eu harvestet die Rechtsinformationssysteme der EU, von Österreich, Holland und UK und verbindet diese Daten mittels Linked (Open) Data Prinzipien zu einer EU-weiten Open Law Plattform. Die akquirierten Daten werden des Weiteren mit Entscheidungen und Rechtspublikationen verlinkt und somit kontextualisiert. Des Weiteren ist eine starke Einbindung der Community geplant. Der Beta Launch ist für Oktober 2015 geplant.

Verwendete Daten: EuroLex (EU: http://eur-lex.europa.eu/homepage.html?locale=de), MetaLex (Linked Legal Data aus Holland): http://doc.metalex.eu/, RIS Österreich: https://www.ris.bka.gv.at/, Legislation.gov.uk (Linked Legal Data UK: http://www.legislation.gov.uk/).

Link, Kontakt: openlaws.eu (10/2015)

(f)Title: reegle.info of REEEP - the information portal on clean energy

Beteiligte Stakeholder: REEEP, SWC (Umsetzer)

Kurzbeschreibung Use Case: Die Länderprofile (Erneuerbare Energie und Energieeffizienz) des Portal reegle.info (siehe: http://www.reegle.info/countries) sind vollständig mittels Linked Open Data Prinzipien und Technologien umgesetzt und bieten damit aktuelle Energie-Länderprofile aller Länder weltweit an. Daten kommen u.a. von REEEP, UNO, WorldBank, Eurostat, Open Energy Information (NREL, USA) und DBpedia (Wikipedia).

Verwendete Daten: Länderinfos DBpedia, Statistiken von UN, Weltbank und Eurostat, Policy Informationen von REEEP und Open Energy Information (US Regierung).

*(i)**Link, Kontakt:***
http://www.reegle.info/countries

(g)Titel: Wolters Kluwer Rechtsinformation in Deutschland

Beteiligte Stakeholder: Wolters Kluwer Deutschland (WKD)

Kurzbeschreibung Use Case: Als Markführer für Rechtspublikationen in Deutschland hat WKD zwei Rechts-Thesauri unter Einbindung von offenen Daten erarbeitet und wiederum als Linked Open Data

publiziert: einen Arbeitsrechts-Thesaurus und einen Gerichts-Thesaurus. Diese werden für die Kategorisierung von Dokumenten und Publikationen verwendet und damit auch für die Integration von WKD-Daten in Drittsystems, wie z.B. Anwaltssysteme.

Verwendete Daten: EuroLex und EuroVoc (http://eurovoc.europa.eu/), Thesaurus Sozialwissenschaften (http://www.gesis.org/unser-angebot/recherchieren/thesauri-und-klassifikationen/thesaurus-sozialwissenschaften/), DBpedia (http://www.dbpedia.org), Standard Thesaurus Wirtschaft (http://zbw.eu/stw/versions/8.08/about.de.html).

*(i)**Link, Kontakt**: http://vocabulary.wolt*

4.3 Weiterführende Literatur

Weiterführende Literatur zu Linked Government (Business) Use Cases:

- David Wood, Linking Government Data, Springer
  Siehe: http://www.springer.com/de/book/9781461417668
- How Linked Data is transforming eGovernment, ISA Programme, EC
  https://drive.google.com/open?id=0B6UD5_Agsb8bRzhFN2ZodEUwb2M
- Study on business models for Linked Open Government Data, European Commission
  https://drive.google.com/open?id=0B6UD5_Agsb8bd0hRQ3pXRDVQVkU
- IEEE - Guest Editors Introduction, Linked Open Government Data
  https://drive.google.com/open?id=0B6UD5_Agsb8bZW4wTUY1cVJkaWs
- Linked Data Platform Use Cases and Requirements
  http://www.w3.org/TR/ldp-ucr/
- HOW LINKED DATA SOLVED A DIGITAL AGE MARKETING PROBLEM
  https://www.programmableweb.com/news/how-linked-data-solved-digital-age-marketing-problem/analysis/2015/08/31

5 Anhang

5.1 Glossar und Abkürzungen

Begriff / Abkürzung	Erläuterung
Abstellung	Ein Zug gilt als abgestellt und somit relevant für die Berechnung des Einsparpotenzials an Strom / CHF. Abgestellt bedeutet: Rollmaterial steht still zwischen zwei fahrplanmässigen Einsätzen als Zug mit einer Zugsnummer.
Closed Data	Daten mit höherer zeitlicher Auflösung und / oder mit zusätzlichen Datenpunkten, die nur gebührenpflichtig in einem geschützten Bereich verfügbar sind
Dynamische Daten	Einzelne Datenpakete, welche zeitnah nach deren Entstehung im Linked Data Store abrufbar sind (Gegenteil: zeitlich verzögert in einem Batch-Lauf importierte Daten)
Entität	Im Sinne eines entity-attribut-value Modells ist eine Entität als Synonym eines RDF-Triple Subjekts zu verstehen.
ETL	Extract, Transform, Load (ETL) ist ein Prozess, bei dem Daten aus mehreren gegebenenfalls unterschiedlich strukturierten Datenquellen in einer Zieldatenbank vereinigt werden.
Graph	Ein Graph ist in der Graphentheorie eine abstrakte Struktur, die eine Menge von Objekten zusammen mit den zwischen diesen Objekten bestehenden Verbindungen repräsentiert. Die mathematischen Abstraktionen der Objekte werden dabei *Knoten* (auch *Ecken*) des Graphen genannt. Die paarweisen Verbindungen zwischen Knoten heißen *Kanten* (manchmal auch *Bögen*). Die Kanten können gerichtet oder ungerichtet sein. Häufig werden Graphen anschaulich gezeichnet, indem die Knoten durch Punkte und die Kanten durch Linien dargestellt werden. (Quelle: Wikipedia). In Zusammenhang mit Linked Data stellen die einzelnen Datenelemente die Knoten und die Links die (gerichteten) Kanten eines Graphen dar. Zwei mit einander über einen Link verbundene Datenelemente werden „Triple“ genannt. In Zusammenhang mit der Speicherung und Verwaltung von Linked Data wird daher von Graph-Datenbanken oder - synonym – von Triple-Stores gesprochen.
LD	Linked Data

Begriff / Abkürzung	Erläuterung
LINDAS	Dieser Linked Data Service schafft ein Netz aus strukturierten Daten, die aus verschiedenen Quellen stammen und zusammen kombiniert weiterverwendet werden können. Er basiert auf semantischen Web-Standards und ist der Auslöser für den Linked Data E-Government Schweiz (LD-eGovCH) Standard.
Linked Data	Linked Data ist ein Konzept, welches die Verknüpfung von Daten über technische und organisatorische Grenzen hinweg ermöglicht. Analog zum World Wide Web, welches beliebige Dokumente (HTML-Seiten) zu einem praktisch grenzenlosen Dokumenten-Raum verknüpft („Web of Documents“), führt Linked Data zu einem globalen Netzwerk unterschiedlichster Datenelemente („Web of Data“). Jedes Datenelement sowie jede Verbindung zwischen zwei Datenelementen (Link) verfügt dabei über einen global eindeutigen und maschinenlesbaren Identifikator (URI).
Linked Government Data	Linked Government Data sind Behördendaten, welche als Linked Data zur Verfügung stehen.
Linked Open Government Data	Linked Open Government Data (LOGD) sind offen zugängliche und frei nutzbare Behördendaten, die als Linked Data zur Verfügung gestellt werden. Der Zugang erfolgt über SPARQL, Download-Services oder API. Die Metadaten zu LOGD werden über OGD-Portale publiziert.
ODBC	Open Database Connectivity, ist eine standardisierte Datenbankschnittstelle welche die Abfragesprache SQL verwendet.
OLAP Würfel	Ein OLAP-Würfel oder Datenwürfel (OLAP cube, ist ein in der Data-Warehouse-Theorie gebräuchlicher Begriff zur logischen Darstellung von Daten. Die Daten werden dabei als Elemente eines mehrdimensionalen Würfels angeordnet. Die Dimensionen des Würfels beschreiben die Daten und erlauben auf einfache Weise den Zugriff. Daten können über eine oder mehrere Achsen des Würfels ausgewählt werden.
Ontologie	Ontologien sind in der Informatik Netze von Hierarchien, in denen Informationen über logische Beziehungen miteinander verknüpft sein können. Konkret meint dies ein formales Modell, welches die Abbildung von Wissen für einen spezifischen Einsatzbereich ermöglicht. Eine Ontologie beschreibt Dinge die existieren (Klassen), Beziehungen zwischen diesen Dingen (Eigenschaften) und den logischen Möglichkeiten der kombinierten Nutzung von Klassen und Eigenschaften (Axiomen).

Begriff / Abkürzung	Erläuterung
Open Data	Kontext MeteoDaten: Zeitlich und / oder regional begrenzte Daten mit grober zeitlicher Auflösung, die gratis und öffentlich zugänglich sind über einen SPARQL Endpoint und eine Web-Oberfläche
Open Government Data (OGD)	Datenbestände des öffentlichen Sektors, die von Staat und Verwaltung im Interesse der Allgemeinheit ohne Einschränkung zur freien Nutzung, zur Weiterverbreitung und zur freien Weiterverwendung offen zugänglich gemacht werden.
Publizieren	Kontext MeteoDaten: Der SPARQL Endpoint für MeteoDaten wird auf LINDAS bekannt gemacht
Relationale Datenbank	Eine relationale Datenbank dient der Verwaltung und persistenten Speicherung von Daten in Computersytemen und beruht auf einem tabellenbasierten relationalen Datenbankmodell.
RDBMS	Relationales Datenbankmanagementsystem welches der Verwaltung von relationalen Datenbanken dient.
RDF	Resource Description Framework. RDF erlaubt die Formulierung von logischen Aussagen über eine Ressource anhand von Subjekt-, Prädikat- und Objekt-Kombinationen. Derartige dreiteilige Aussagen werden als Tripel (englisch: Triples) bezeichnet.
RDF Store	Auch Triplestore genannt, in welchem Daten direkt im RDF Format gespeichert werden können.
REST	Representational State Transfer REST bezeichnet ein Programmierparadigma für verteilte Systeme, insbesondere für Webservices. REST ist eine Abstraktion der Struktur und des Verhaltens des World Wide Web. REST hat das Ziel, einen Architekturstil zu schaffen, der die Anforderungen des modernen Web besser darstellt.
Ressource	Im RDF-Kontext kann eine Ressource alles sein, was durch einen RDF-Graphen beschrieben ist. Ressourcen können durch URIs adressiert werden.
Semantic Web	„Das Semantic Web erweitert das Web, um Daten zwischen Rechnern einfacher austauschbar und für sie einfacher verwertbar zu machen; so kann beispielsweise der Begriff „Bremen“ in einem Webdokument um die Information ergänzt werden, ob hier ein Schiffs-, Familien- oder der Stadtname gemeint ist. Diese zusätzlichen Informationen explizieren die sonst nur unstrukturiert vorkommenden Daten. Zur Realisierung dienen Standards zur Veröffentlichung und Nutzung maschinenlesbarer Daten, insbesondere RDF.“ (nach Wikipedia)

Begriff / Abkürzung	Erläuterung
Serialisierung	Abbildung strukturierter Daten in in einer definierten Syntax in welcher die Daten ausgetauscht werden können
SPARQL	Abfragesprache für in RDF gespeicherte Informationsressourcen.
SQL	SQL ist eine Datenbanksprache zur Definition von Datenstrukturen in relationalen Datenbanken sowie zum Bearbeiten (Einfügen, Verändern, Löschen) und Abfragen von darauf basierenden Datenbeständen.
Subjekt	Erster Teil eines RDF-Triples, gibt an auf wen oder was sich das Statement bezieht.
Triple, Triple Store	Siehe RDF und Graph
URI	Eindeutiger Bezeichner einer Ressource.
URL	Ist eine via HTTP erreichbare Adresse einer Ressource
Vokabular	Sammlung von „Begriffen“ für einen bestimmten Zweck. Vokabulare spielen im Kontext von Linked Data eine wichtige Rolle, speziell für die Datenintegration. Die Nutzung des Begriffs „Vokabular“ überlappt sich mit dem Begriff Ontologie.

5.2 Verzeichnisse

Abbildungen

Abbildung 5: Semantic Web Layer Cake 5
Abbildung 6: Das Fünf Sterne Modell des Datenpublizierens 9
Abbildung 7: Das „Web of Data“ – die Linked Data Cloud. Eine interaktive Version findet sich auf http://lod-cloud.net/ 11
Abbildung 8: Process Linked Data 12
Abbildung 9: Consume Linked Data 15
Abbildung 10: Eingebettete Metadaten in der Linked Data Version von Biel/Bienne bei Swisstopo 16
Abbildung 11: Automatische optimale Darstellung von Rohdaten zu Interlaken aus Wikipedia/DBpedia 17
Abbildung 12: UniProt Datensatz Statistik 2017, Quelle: https://twitter.com/uniprot/status/864770108455092224 19
Abbildung 13: JVM & System Memory Empfehlungen vom Stardog Triplestore. Quelle: http://www.stardog.com/docs/#_memory_usage 20
Abbildung 14: Beispiel eines generischen RDF Data Cube Browsers, angewendet auf BAFU Umwelt Basisdaten 23
Abbildung 15: RAWGraphs Auswahl der möglichen Visualisierungen 24
Abbildung 16: Auswahl der Dimensionen der Daten in RAWGraphs 24
Abbildung 17: Wetterdaten ausgewählt auf Basis einer Auswahl in Tableau (Quelle: https://www.tableau.com/products/desktop) 25
Abbildung 18: SPARQL Abfrage der aufgehobenen Gemeinde im Jahr 2016 mit entsprechender Visualisierung der Gemeinde. Die Umrisse der Gemeinde können direkt in der SPARQL-Konsole dargestellt werden, in diesem Beispiel aufbauend auf Openstreetmap Karten. Für Swisstopo wurde ein Plugin entwickelt, welches die Schweizer Karten in diese Sicht Integriert. 26
Abbildung 4: Konzeptionelle Sicht auf den Show Case 29
Abbildung 20: Tagesansicht 30
Abbildung 21: Zusammenfassung mit kumulierten Einsparungen 31

Tabellenverzeichnis

Tabelle 1: Fünf Sterne Modell 10
Tabelle 2: Abgrenzung zu anderen Konzepten 18
Tabelle 3: Rollen 28
Tabelle 6: Linked Data Projekte Schweiz 35