Das Linked Open Data Warehouse

Richard Figura[1] and Alexander Willner[2]

[1] CISS TDI GmbH
Sinzig, Germany
r.figura@ciss.de
[2] Fraunhofer FOKUS / Technische Universität Berlin
Berlin, Germany
alexander.willner@{fokus.fraunhofer.de|tu-berlin.de}

Zusammenfassung Daten sind einer der wichtigsten Rohstoffe der digitalen Welt, mit wachsender wirtschaftlicher Bedeutung, für Unternehmen und öffentliche Verwaltung. Um dieses Potential auszuschöpfen, ist es von entscheidender Bedeutung, dass heterogene Daten unterschiedlichster Quellen miteinander verknüpft und kombiniert werden können. Ein vielversprechender Lösungsansatz hierzu sind semantische Technologien, die vom World Wide Web Consortium (W3C) standardisiert werden. Eine große Herausforderung besteht jedoch darin, wie sich dieses Konzept in bereits etablierte Prozesse und bestehende Infrastruktur integrieren lässt. In dieser Arbeit wird ein Ansatz vorgestellt, wie Informationen verschiedenster Quellen durch ein Data Warehouse, basierend auf Oracle Spatial und Graph, verarbeitet werden können. Das Data Warehouse erlaubt die Wiederverwendung bereits etablierte Technologien, beispielsweise des Open Geospatial Consortium (OGC), um heterogene Daten verschiedener offener-, sowie privater Quellen miteinander zu kombinieren. Datenlieferanten können somit eigene Daten vorverarbeiten und per Dienst, oder mit Hilfe semantischer Technologien Verfügung stellen. Einem Datenkonsumenten hingegen erlaubt es die einfache Integration offener Daten unterschiedlichster Quellen in die eigene Datenlandschaft.

Keywords: Linked Open Data, Semantic Web, Data Warehouse

1 Motivation

In der heutigen Wissensgesellschaft ist die Information die wichtigste Ressource des 21. Jahrhunderts. Eine wichtige Quelle von Informationen sind Daten aus dem öffentlichen Bestand wie Angaben zur Bevölkerung, Bildung und Wissenschaft, Geobasisdaten, Gesetze, Gesundheit, Infrastruktur, Tourismus, Politik, Transport und Verkehr, Umwelt oder Wirtschaft. Sind diese Daten öffentlich zugänglich, so unterstützt dies mehr Transparenz, führt zu mehr Teilhabe seitens der Bevölkerung und trägt zu mehr Innovation bei. So wird laut einer Studie[6], der indirekte und direkte Nutzen offener Daten alleine in Europa auf 140 Milliarden Euro pro Jahr geschätzt, andere Studien[5] gehen sogar von einem jährlichen Mehrwert von bis zu 900 Milliarden Euro aus.

Ein entscheidender Mehrwert wird darüber hinaus dadurch gebildet, wenn unterschiedliche Daten miteinander in Bezug gesetzt werden. Eingängige Beispiele sind die Nutzung von Stauinformationen in der Navigation oder die Visualisierung von Umweltdaten auf einer geographischen Karte.

Aus dieser Verknüpfung verschiedenster Daten aus unterschiedlichen Quellen resultieren jedoch auch einige Herausforderungen. Daten lediglich maschinenlesbar öffentlich verfügbar zu machen, erlaubt nur begrenzt eine inhaltlich kohärente Verknüpfung verschiedenster, teilweise inkompatibler Daten miteinander. Beispielsweise können Geodaten in unterschiedlichsten Datenformaten kodiert sein, verschiedene Kartenprojektionen nutzen und Daten in verschiedener Qualität oder für unterschiedliche Gebiete vorhalten.

Hürden, die die Nutzung der offenen Daten für Dritte nachhaltig erschweren, entstehen im Allgemeinen daraus, dass Daten unter anderem

- aus heterogenen Datenquellen kommen, beispielsweise durch aus dem Dateisystem, einem Web Feature Service (WFS) oder per File Transfer Protocol (FTP);
- heterogene Datenmodelle nutzen, beispielsweise als Comma Separated Values (CSV), als Extensible Markup Language (XML) oder im JavaScript Object Notation (JSON) Format;
- nur Fragmente der benötigten Daten enthalten, beispielsweise unterschiedliche Ländergrenzen;
- unterschiedliche Datenqualität aufweisen, beispielsweise redaktionell aufbereitet.

Daraus resultiert, dass Daten adressatengerecht aufbereitet und Möglichkeiten geschaffen werden sollten, um Informationen verschiedener Quellen miteinander zu verknüpfen.

2 Grundlagen

Ziel des Ansatzes ist es, existierende Daten öffentlichen und nicht-öffentlichen Datenquellen in ein System zu importieren, zu konvertieren und zu verknüpfen, um diese einheitlich über die Schnittstellen und Datenformate bereitzustellen, die in den jeweiligen Anwendungsdomänen üblich sind. Im Kontext der Geodatenverarbeitung kommen dafür Geodata Warehouses (GDWs) zum Einsatz und beispielsweise sind hier Zugriffstechnologien üblich, die Konform zum Open Geospatial Consortium (OGC) sind. Damit werden weitere Nutzungsmöglichkeiten der Daten geschaffen und existierende Barrieren weiter gesenkt.

Über den Anwendungsbereich der Geodaten hinaus, hat sich in den letzten Jahren ein breites Ökosystem aus offenen verknüpften Daten gebildet, in das zum einen Datenschätze gehoben als auch integriert werden können. Unter dem Begriff Linked Open Data (LOD) [1] werden Konzepte verstanden, um Daten leichter auffindbar und verständlich, wiederverwendbar, mit existierenden Werkzeugen verarbeitbar, mit anderen Daten kombinierbar zu machen. Der Wert Offener Daten erhöht sich durch diese Herangehensweise signifikant, für Unternehmen und

Gesellschaft. Grundlage ist ein Konzept das bereits 2006 entworfen wurde, um weltweit verteilte und heterogene Daten miteinander verknüpfen zu können [2]. Die Technologien um das so genannte Semantic Web [3] wurden im Rahmen des World Wide Web Consortium (W3C) standardisiert.

Für die Beurteilung der Qualität veröffentlichter Daten kann das so genannte 5-Sterne-Modell herangezogen werden. Wie in Abbildung 1 dargestellt gibt es fünf Untergliederungen.

1. Daten sind öffentlich im Web verfügbar, beispielsweise im Portable Document Format (PDF) Format;
2. Daten sind strukturiert maschinenlesbar, beispielsweise im Microsoft Excel (XLS) Format;
3. Daten liegen in einem freien Format vor, beispielsweise im CSV Format;
4. Daten nutzen offene Standards zur einheitlichen Identifizierung, beispielsweise im Resource Description Framework (RDF) [4] Format;
5. Daten sind verknüpft mit anderen externen Daten.

Nach dieser Definition fallen viele Daten, die heutzutage unter dem Begriff Öpen Data"veröffentlicht werden, in die zweite oder dritte Kategorie. Zwar wird dadurch einer wichtige Grundlage geschaffen, um Daten offen bereitzustellen und nutzbar zu machen. Um die umfangreichen und weiter wachsenden Datenbestände jedoch intelligent zu vernetzen und damit Mehrwerte zu generieren, bedarf es weiterer Arbeiten.

Abbildung 1: 5-Sterne-Modell für offene, verknüpfte Daten ([2])

3 Eigener Ansatz

Der eigene Ansatz basiert im Kern auf Oracle Spatial and Graph, um RDF-Graphen innerhalb eines relationales Datenbankmanagementsystems von Oracle zu speichern. Hierbei werden sowohl existierende SQL-basierte Zugriffe mit SPARQL-basierten Abfragen kombiniert.

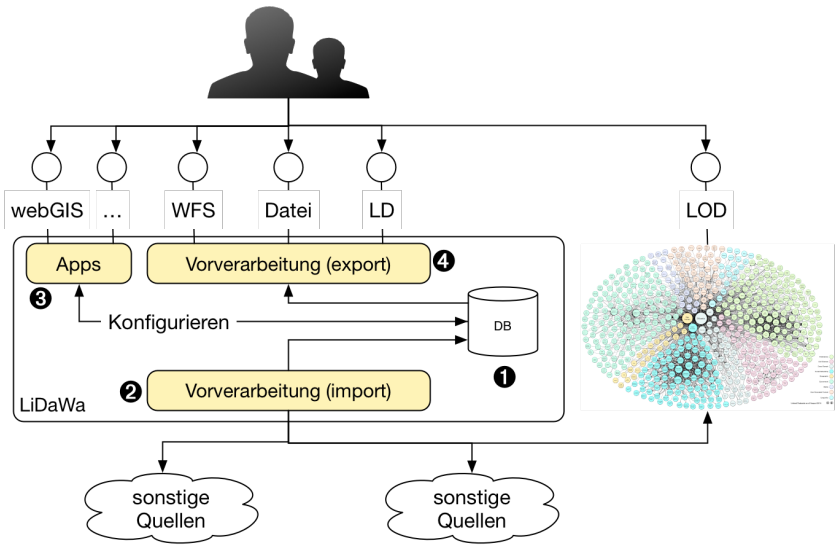

Abbildung 2: Architektur des eigenen Ansatzes.

Wie in Abbildung Abbildung 2 dargestellt, werden die erforderlichen Funktionen von den folgenden vier Komponenten des Linked Open Data Warehouse (LODW) bereitgestellt:

1. **Datenbank**: Die Datenbank speichert die importierten Daten und deren Metadaten in einem speziellen Datenmodell. Hierzu ergänzt sie klassische Datenbanken durch die Verwaltung von Semantiken. Dadurch ist es möglich, räumliche Abfragen, oder Verschneidungen direkt auf der Datenbank auszuführen und Metainformationen zu den gespeicherten Daten abzufragen.
2. **Import**: Beim Import werden Daten zur Datenbank hinzugefügt, sodass diese für einen Nutzer verfügbar sind und in nachfolgenden Arbeitsschritten verarbeitet, bzw. veredelt werden können. Diese Komponente stellt zudem sicher, dass die Daten aus verschiedenen Quellen einlesbar sind, entsprechend aufbereitet- und durch semantischen Informationen ergänzt werden. Die Datenaufbereitung schließt hierbei eine Konvertierung der Daten ins gewünschte Zielformat, sowie eine Homogenisierung der Daten mit ein.

3. **Apps**: Das Application Center bietet die Möglichkeit, gespeicherte Daten aus der Datenbank zu attributieren. Zudem erlaubt es, die attributieren Daten innerhalb eines Web-GIS anzuzeigen und bietet Routinen, um eine Applikation für die attributieren Daten zu definieren. Applikationen können unter anderem definieren, in welchen Formaten angezeigte Daten exportiert werden können, oder ob ein Ausschneiden der angezeigten Daten möglich ist.
4. **Export**: Der Export erlaubt die Veröffentlichung der gespeicherten Roh- oder veredelten Daten. Die Daten können hierfür in das gewünschte Zielformat konvertiert werden und dann über die folgenden drei Wege veröffentlicht werden: Als LOD innerhalb des semantischen Webs, als Webdienst WFS oder als Dateiexport.

Das Ergebnis ist ein System mit folgenden Eigenschaften:

- Importmöglichkeit aller klassischen Datenquellen, sowie gängiger Formate. Dies betrifft insbesondere eigene Daten und öffentliche Daten.
- Importmöglichkeit aller Daten aus dem semantischen Web und dadurch eine Erweiterung des Nutzbaren Datenportfolio.
- Standardisierter Zugriff über eine Datenbank und die damit verbundene Möglichkeit, importierte Daten mit Standardwerkzeugen zu verarbeiten.
- vereinfachte Visualisierung durch das Application Center.
- Vereinfachte Definition von Web-Apps durch das Application Center
- Exportmöglichkeit der gespeicherten Roh- oder veredelten Daten in alle gängige Formate.
- Exportmöglichkeit der gespeicherten Roh- oder veredelten Daten in das semantische Web.

4 Geschäftsmodelle

Bei der Benennung der Vorteile des LODW muss in die folgenden zwei Anwendergruppen unterschieden werden: Ein *Bereitsteller* von Daten möchte eigene Daten der Öffentlichkeit zur Verfügung stellen, damit diese einen Mehrwert für Gesellschaft und Wirtschaft erzeugen. Bei dieser Anwendergruppe handelt es sich häufig öffentliche Einrichtungen, die politischen Vorgaben folgen. Ein *Anwender* auf der anderen Seite möchte die zur Verfügung gestellten Daten verwerten und in die eigenen Datenlandschaft integrieren. Durch die Kombination von frei verfügbaren mit öffentlichen Daten wird ein Mehrwert erzeugt, der neue Dienstleistungen und Produkte erlaubt. Bei dieser Anwendergruppe handelt es sich vornehmlich um Unternehmen aus der freien Marktwirtschaft.

Diese beiden Zielgruppen können jeweils vom Betrieb eines LODW profitieren.

Für Bereitsteller von Daten bietet das LODW die Möglichkeit eigene Daten direkt im Semantischen Web zur Verfügung zu stellen. Durch die Verwendung zukunftssicherer Technologien sind wird im Vergleich zu möglichen Alternativen die größte und nachhaltigste Wirkung für Gesellschaft und Wirtschaft erreicht. Zudem ermöglicht das LODW eine Qualitätsprüfung der Daten, bereits bei deren Bereitstellung. Deren Ergebnis kann zum einen helfen, die Qualität der

Daten schrittweise zu verbessern, zum anderen gibt das Ergebnis dieser Prüfung wichtige Hinweise für den Anwender der bereitgestellten Daten. Zuletzt können die Informationen aus dem LODW auch auf traditionellem Wege bereitgestellt werden, beispielsweise als Datei-Export oder als Dienst. Bereits etablierte Wege einer Datenbereitstellung stehen im LODW also weiterhin zur Verfügung.

Für den Anwender von Daten bietet das LODW eine zentrale Datenhaltung, durch die ein einheitlicher Zugriff auf die unternehmensweiten Daten erlaubt wird. Durch eine Schnittstelle zum Semantischen Web wird es möglich, zusätzliche - bereits frei verfügbare - Informationen einfach mit eigenen Daten zu kombinieren. Neben einer vereinfachten Integration von Daten verschiedener Quellen bedeutet dies zudem eine Verbesserung der Datenqualität, da stets auf die aktuellen Daten er entsprechenden Datenquelle zugegriffen werden kann, ohne Kopien dieser erzeugen zu müssen. Insgesamt hilft das LODW hierdurch die Kosten für Datenhaltung, Pflege und Integration zu senken

Danksagung

Diese Arbeit wurde teilweise durch das Limbo-Projekt (Nr. 19F2029I) des Bundesministeriums für Verkehr und digitale Infrastruktur (BMVI) unterstützt.

Kontaktdaten

Richard Figura
CISS TDI GmbH
Barbarossastraße 36
D-53489 Sinzig
Telefon: +49 (0) 2642 9780-11
Fax: +49 (0) 2642 9780-10
E-Mail r.figura@ciss.de
Internet: http://www.ciss.de

Dr. Alexander Willner
Fraunhofer FOKUS / TU Berlin
Marchstr. 23
D-10587 Berlin
Telefon: +49 (0) 30 3463 7116
Fax: +49 (0) 3463 997116
E-Mail: alexander.willner@tu-berlin.de
Internet: http://av.tu-berlin.de/willner

Literatur

[1] Florian Bauer und Martin Kaltenböck. „Linked open data: The essentials“. In: *Edition mono/monochrom, Vienna* (2011).
[2] Tim Berners-Lee. *Linked Data.* World Wide Web Design Issues. Juli 2006.
[3] Tim Berners-Lee, James Hendler und Ora Lassila. „The Semantic Web“. In: *Scientific American* 284.5 (Mai 2001), S. 34–43.
[4] Richard Cyganiak, David Wood und Markus Lanthaler. *Resource Description Framework (RDF) 1.1 Concepts and Abstract Syntax.* Recommendation. World Wide Web Consortium (W3C), Feb. 2014.
[5] James Manyika. *Open data: Unlocking innovation and performance with liquid information.* McKinsey, 2013.
[6] Graham Vickery. „Review of recent studies on PSI re-use and related market developments“. In: *Information Economics, Paris* (2011).