X.media.press

Tassilo Pellegrini
Harald Sack
Sören Auer (Hrsg.)

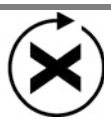
X.media.press ist eine praxisorientierte Reihe zur Gestaltung und Produktion von Multimedia-Projekten sowie von Digital- und Printmedien.

Linked Enterprise Data

Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien

Springer Vieweg

X.media.press



X.media.press ist eine praxisorientierte Reihe zur Gestaltung und Produktion von Multimedia-Projekten sowie von Digital- und Printmedien.

Tassilo Pellegrini · Harald Sack · Sören Auer
Herausgeber

Linked Enterprise Data

Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien

Springer

Herausgeber

Tassilo Pellegrini
Institut für Medienwirtschaft
Fachhochschule St. Pölten
St. Pölten, Österreich

Sören Auer
Institut für Informatik III
Rheinische Friedrich-Wilhelms-Univ. Bonn
Bonn, Deutschland

Harald Sack
Hasso-Plattner-Institut für
Softwaresystemtechnik GmbH
Universität Potsdam
Potsdam, Deutschland

ISSN 1439-3107
ISBN 978-3-642-30273-2 ISBN 978-3-642-30274-9 (eBook)
DOI 10.1007/978-3-642-30274-9

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Springer ist Teil der Fachverlagsgruppe Springer Science+Business Media
www.springer.com

Vorwort der Herausgeber

Das World Wide Web ist im Begriff sich von einer weltweiten Sammlung vernetzter Dokumente hin zu einem Netzwerk verknüpfter Daten zu entwickeln. Dieses „Web of Data“ erlebt seit einigen Jahren ein immenses Wachstum. Es speist sich aus unzähligen Datenquellen unterschiedlichster Größe, Qualität und Themen, die entweder offen oder geschlossen zur Verfügung stehen und bereits jetzt vielerorts eine wichtige Komponente in der betrieblichen Datenverarbeitung darstellen. In Kombination mit dem weithin bekannten „Web of Documents“ bildet das „Web of Data“ ein neues Öko-System und eine grundlegende Infrastruktur für Software-Anwendungen und Dienste der Zukunft. Prominente Entwicklungen wie etwa „Big Data“, Open (Government) Data, Cloud Computing und Service-Orientierung sind Teilaspekte eines weitreichenden Wandels im Enterprise Data Management, in dessen Zentrum die Nutzung und Bewirtschaftung verteilter Daten steht.

Doch die stetig wachsende Verfügbarkeit qualitativ hochwertiger und strukturierter Daten sowohl innerhalb als auch außerhalb von Unternehmen veranlasst die Frage nach neuen Methoden und Technologien des Enterprise Data Managements. Konventionelle Datenbereitstellungsstrategien in Form von (semi-)strukturierten Dokumenten (z. B. HTML, CSV-Dateien) oder proprietären Programmierschnittstellen werden nur bedingt den Ansprüchen hoch vernetzter und dynamischer Daten-Ökosysteme gerecht. Mit jeder zusätzlichen Quelle steigen die Integrationsaufwände, Veränderungen in der Datenbankstruktur gehen oftmals zu Lasten der Systemintegrität und Aktualisierungen der Datenbasis sind nur mit hohem Aufwand in Echtzeit verfügbar.

Hier setzt der „Linked Data“ Ansatz an, der eine höchst-mögliche Flexibilität und technische Interoperabilität in der unternehmerischen Datenhaltung anstrebt und so die kosteneffiziente und zeitkritische Integrierbarkeit, eindeutige Interpretierbarkeit und Wiederverwendbarkeit von Daten ermöglicht. Linked Data bedient sich dazu sogenannter Semantic Web Standards, um existierende Datenbestände hoch strukturiert aufzubereiten und plattformunabhängig für die weitere Integration und Syndizierung bereitzustellen.

Mit diesem Band thematisieren die Herausgeber die Bedeutung neuer Formen des technologisch gestützten (Meta-)Daten-Managements für die voranschreitende Vernetzung und Integration verteilter, heterogener Datenbestände zur Unterstützung des betrieblichen Informations- und Wissensmanagements. Hierbei spielt vor allem der Einsatz von Se-

mantic Web Technologien, sowohl als Produktions- als auch als Distributionsinfrastruktur für umfassende Datensammlungen, eine zentrale Rolle. Denn durch Semantic Web Technologien werden Daten zu Netzgütern und erlauben neue Formen der Datenhaltung und Bewirtschaftung. Hierbei stellt sich die Frage, inwieweit die föderalen, selbstorganisierenden und kollaborativen Mechanismen des Webs in den kontrollierten Umgebungen einer Organisation sinnvoll zum Einsatz gebracht werden können, um neue Ressourcen aus bestehenden Informationsbeständen zu generieren und um von der Fülle an verfügbaren, qualitativ hochwertigen (offenen) Datenquellen im Web zu profitieren. Dazu diskutieren die einzelnen Beiträge technologische und methodische Aspekte des semantisch gestützten Datenmanagements und zeigen mittels Fallstudien existierender Implementierungen, wie eine gesteigerte Konnektivität und Interoperabilität von Datenquellen das Enterprise Data Management beeinflussen (werden).

Der Band „Linked Enterprise Data“ richtet sich an eine technisch versierte Leserschaft, die sich mit den Grundlagen und Anwendungsmöglichkeiten des Linked Data Prinzips für das betriebliche Informations- und Wissensmanagement vertraut machen möchte. Insbesondere adressiert der Band existierende und in Ausbildung befindliche Professionalisten, die als CTOs, CIOs, Enterprise Architects, Projektmanager und Applikationsentwickler in Unternehmen, Non-Profit Organisationen oder öffentlichen Einrichtungen arbeiten und mit Fragen der Skalierbarkeit, Flexibilität, Robustheit und Nachhaltigkeit von Informationssystemen befasst sind.

Der Band gliedert sich in drei Abschnitte. Der erste Abschnitt gliedert sich in Kap. 1 bis 3 und erläutert die technologischen Grundlagen und die geschäftsmodellrelevanten Aspekte des betrieblichen Einsatzes von Linked Data. Der zweite Abschnitt mit den Kap. 4 bis 9 diskutiert methodische Aspekte von Linked Data Technologien zur Lösung konkreter Probleme sowie technologische Entwicklungsmöglichkeiten im betrieblichen Datenmanagement. Der dritte Abschnitt mit den Kap. 10 bis 14 rundet die theoretischen Erläuterungen mit Fallstudien von konkreten Implementierungen von Linked Data Technologien ab.

Im Folgenden werden die Beiträge kurz vorgestellt.

In Kap. 1 erläutert Andreas Blumauer die Prinzipien von Linked Data in der Aggregation, Verwaltung und Bewirtschaftung von geschäftsrelevanten Datenquellen. Er zeigt, dass Linked Data basierte Datenmodelle weniger abstrakt sind als XML-Schemata oder relationale Datenbankmodelle und deshalb besser geeignet sind, Informationen für Mensch und Maschine verfügbar zu machen. Sogenannte „semantische Wissensgraphen“ oder Ontologien können hierbei modular und inkrementell entwickelt werden und mit den Geschäftsanforderungen flexibel mitwachsen. Linked Data Graphen tragen direkt zur Verbesserung der User-Experience bei und generieren Netzeffekte rund um interoperable Datenbestände. Der Beitrag führt diese Aspekte weiter aus und diskutiert vier Anwendungsfälle für den praktischen Einsatz von Linked Data im Unternehmen.

Kapitel 2 von Harald Sack bietet einen grundlegenden Überblick über das Thema Linked Data und führt in die dazugehörigen Basistechnologien ein. Nach der detaillierten

Erläuterung der Bedeutung und Funktion der eindeutigen Identifikation von Ressourcen in Wissensbasen wird in das Resource Description Framework (RDF) zur einfachen Modellierung von Fakten eingeführt. Linked Data lebt von der Verknüpfung der Fakten untereinander sowie mit zugrundeliegenden Wissensrepräsentationen in Form von Ontologien. In diesem Zusammenhang werden auch RDF(S) und OWL als formale Ontologiebeschreibungssprachen vorgestellt, um Möglichkeiten und Grenzen des Ansatzes aufzuzeigen. Weiterführend werden Möglichkeiten zur Nutzung von Linked Data in unternehmerischen Anwendungen vorgestellt sowie auf die Veröffentlichung eigener Datensätze als Linked Data eingegangen.

In Kap. 3 diskutiert Tassilo Pellegrini vor allem rechtliche Aspekte der Bewirtschaftung von vernetzten Daten entlang der Content Value Chain. Dies umfasst zum einen die Integration und Verwendung externer Daten im Zuge der Content-Verarbeitung, zum anderen die Wahl des richtigen Lizenzmodells für die Veröffentlichung eigener Daten als Linked Open Data. Ausgehend von unterschiedlichen Asset-Typen, die bei der Generierung von Linked Data anfallen, zeigt der Beitrag, welche Asset-Typen durch welches Rechtsinstrument geschützt werden können. Ein besonderes Augenmerk liegt auf der Kombination offener und geschlossener Lizenzinstrumente zu Zwecken der Diversifikation von Geschäftsmodellen.

In Kap. 4 gehen Sören Auer, Jörg Unbehauen und Rene Pietsch auf methodische Probleme der Integration verteilt vorliegender Unternehmensdaten ein. Sie argumentieren, dass Daten-Intranets auf Basis von Linked Data Technologien die existierenden Intranet- und SOA-Landschaften in großen Unternehmen erweitern und flexibilisieren. Hierbei bietet Linked Data die Möglichkeit der Nutzung von Daten aus der inzwischen auf über 50 Mrd. Fakten angewachsenen Linked Open Data (LOD) Cloud. Im Ergebnis kann ein unternehmensinternes Daten-Intranet, das sowohl interne als auch externe Quellen integriert, dazu beitragen die Brücke zwischen strukturiertem Datenmanagement (in ERP, CRM, SCM Systemen) sowie semi- und unstrukturierten Informationen (Dokumente, Wikis, Portale) der Intranet-Suche zu schlagen.

Diese Ausführungen werden durch Robert Isele in Kap. 5 vertieft. Sein Beitrag behandelt die notwendigen Prozesse, um eine globale Sicht auf mehrere Datenquellen herzustellen, sodass diese für eine gemeinsame Abfrage zur Verfügung stehen. Im Kern steht das Problem, dass Linked Data Publisher eine Vielzahl verschiedener Vokabulare verwenden um Informationen zu repräsentieren. Es gilt zunächst die Datensets in ein konsistentes Zielvokabular überzuführen und in einem zweiten Schritt, Ressourcen in unterschiedlichen Datensets, welche dasselbe Realwelt-Objekt repräsentieren, zu identifizieren und zu verknüpfen.

In Kap. 6 illustrieren Philipp Frischmuth, Michael Martin, Sebastian Tramp und Sören Auer am Beispiel der Anwendung OntoWiki aktuelle Ansätze in der Linked Data-Visualisierung und diskutieren deren Bedeutung im Enterprise Information Management. Die visuelle Aufbereitung von Linked Data sowohl für Zwecke der Prozessverarbeitung als auch zur Konsumierung durch Endanwender ist ein wichtiges Designelement in der

unternehmerischen Aneignung von Semantic Web Technologien, insbesondere im Zuge der Kuratierung und Qualitätssicherung verteilter Daten.

Philipp Cimiano und Christina Unger diskutieren in Kap. 7 das Problem der Multilingualität in verteilten Wissensbasen, die über Länder- und Sprachgrenzen hinweg erzeugt und genutzt werden. Die Autoren besprechen Verfahren, mit denen Datenschemata, die für verschiedene Länder entwickelt wurden, synchronisiert werden können, um die Aggregation und Integration von Daten über Länder und Sprachgrenzen hinweg zu ermöglichen. Darüber hinaus erläutern sie, wie Linked Data mit linguistischen Informationen angereichert werden kann, und betrachten einige Anwendungen, die zeigen, wie solche Informationen für die Generierung und die Interpretation natürlicher Sprache verwendet werden können.

In Kap. 8 beschäftigen sich Sebastian Bayerl und Michael Granitzer mit dem Einsatz von Linked Data Technologien im Data-Warehousing. Data-Warehousing bezeichnet die technologische Realisierung analytischer Datenbestände sowie entsprechender Schnittstellen zu deren Exploration und Analyse. Linked Data bietet vor allem mit der vor Kurzem begonnenen Entwicklung des RDF Data Cube Vokabulars neue Entwicklungsmöglichkeiten für Data-Warehousing Technologien und deren Einsatzspektrum. Der Beitrag stellt die Grundlagen zu Data-Warehouses vor und führt in das RDF Data Cube Vokabular als Linked Data Äquivalent ein. Beide Grundlagen dienen der Diskussion sowohl der Anwendung von RDF Data Cubes im Data-Warehousing als auch der Erweiterung traditioneller Data-Warehousing Ansätze, z. B. durch Integration offener Daten in Data-Warehousing Prozessen.

Kapitel 9 bietet einen kompakten Einstieg in das Thema Reasoning auf Basis strukturierter Daten zu Zwecken der automatischen Erschließung neuen Wissens aus oder der Qualitätssicherung von Datenbeständen. Dazu beschreiben Jens Lehmann und Lorenz Bühmann die Grundlagen des Reasonings in RDF/OWL-Wissensbasen und besprechen unterschiedliche Methoden des Reasonings. Weiters gehen sie auf Herausforderungen und Grenzen des Einsatzes von Reasoning-Technologien im Kontext von Linked Data ein.

In Kap. 10 beschreibt Anja Jentzsch die Linking Open Data Cloud, eine umfangreiche Sammlung von offen lizenzierten, vernetzen Daten und Kristallisationspunkt des Web of Data. Diese Data Cloud besteht aus mittlerweile 82 Milliarden RDF-Tripeln verteilt auf fast 1000 Datensätze, die vielfältige thematische Domänen abdecken. Der Beitrag analysiert ausgewählte Datensätze, welche im gemeinschaftlich gepflegten LOD Cloud Data Catalog eingetragen sind, und illustriert, wie diese Datensätze und deren Verlinkungen über die Linking Open Data Cloud visualisiert werden.

In Kap. 11 erläutern Natalja Friesen und Christoph Lange die Implementierung von Linked Data im Kontext Digitaler Bibliotheken. Wichtige Ziele bei der Entwicklung Digitaler Bibliotheken sind Informationen leicht auffindbar zu machen, sie miteinander zu verknüpfen, sowie die Inhalte der Bibliothek für Mensch und Maschine nutzbar zu machen. Dazu stellen die Autoren wichtige Standards, Vokabulare und Ontologien für bibliographische (Meta-)Daten vor und diskutieren Herausforderungen beim Publizieren Digitaler

Bibliotheken als Linked Data. Zu den Herausforderungen gehören Datenmodellierung, Mapping, sowie Verknüpfung der Daten miteinander und mit anderen Datenbeständen. Als konkrete Anwendungsfälle geben die Autoren einen Überblick über die Europeana und die Deutsche Digitale Bibliothek (DDB), stellen aber auch weitere Digitale Bibliotheken vor, die Linked Data einsetzen.

In Kap. 12 erläutern Michael Gorriz und Kai Holzweißig den Einsatz von Linked Data Technologien bei einem großen deutschen Autohersteller. Laut ihrer Argumentation erlauben Unternehmen in der Automobilindustrie gegenwärtig einen tiefgreifenden Wandel. Informationstechnologie bestimmt immer mehr die Art und Weise, wie Unternehmen arbeiten, und insbesondere die Entstehungsprozesse ihrer Produkte und Dienstleistungen. Kurzum: Die Idee des digitalen Unternehmens ist auch heute schon in traditionell geprägten Industriezweigen wie der Automobilindustrie zur Wirklichkeit geworden. Aufgrund der verschiedenen technologischen, organisationalen und kulturellen Herausforderungen, die dieser Paradigmenwechsel bedingt, bedarf es neuer Konzepte und Technologien, um diesen Wandel nachhaltig zu unterstützen. Im Rahmen des vorliegenden Artikels wird aufgezeigt, dass die Idee von Linked Data ein solches Konzept darstellen kann. Neben einer kurzen Diskussion der Grundlagen wird detailliert aufgezeigt, welche Anwendungsfälle und Mehrwerte für eine Praxisanwendung von Linked Data existieren.

In Kap. 13 diskutieren Harald Sack und Jörg Waitelonis Einsatz von Linked Data zur Verbesserung der Auffindbarkeit audio-visueller Information. Insbesondere Videodaten sind auf dem besten Wege zur bedeutendsten Informationsquelle im World Wide Web zu werden. Bereits heute werden pro Minute mehr als 100 Stunden Videomaterial von den Benutzern auf Videoplattformen wie YouTube eingestellt. Bei dieser gewaltigen Menge an unstrukturierten multimedialen Daten wird auch die gezielte Informationssuche immer schwieriger, da eine inhaltsbasierte Suche mit Hilfe von textbasierten Metadaten realisiert wird, die entweder manuell oder mittels unzuverlässiger automatischer Analyseverfahren gewonnen werden. Hier bietet die semantische Videosuche einen Ausweg, die aufbauend auf einer Vielzahl unterschiedlicher Analyseverfahren versucht, textbasierte Metadaten inhaltlich miteinander in Bezug zu setzen und zielsicher die gewünschten Ergebnisse zu finden. Darüber hinaus ermöglicht es den zu Grunde liegenden Suchraum, d. h. das gesamte Videoarchiv ähnlich dem Stöbern in einem gutsortierten Bücherregal zielstrebig zu durchmustern und auf diese Weise hilfreiche neue Informationen zu finden. Die Videosuchmaschine yovisto.com implementiert zahlreiche visuelle Analyseverfahren und kombiniert diese prototypisch in einer explorativen semantischen Suche.

Kapitel 14 beschließt den Band mit einer kompakten Darstellung des Einsatzes von Linked Data beim deutschen Fachverlag Wolters Kluwer Deutschland. Christian Dirschl und Katja Eck zeigen anhand von Businessanforderungen, wie sich die Wertschöpfungskette innerhalb eines Medienhauses unter Einbeziehung von Linked Data weiterentwickeln kann. Insbesondere die systematische Trennung von textlichem Content und Metadaten eröffnet völlig neue Möglichkeiten im Gesamtprozess. Die strukturierte Einbindung externer Wissensquellen stellt dabei ein nicht zu vernachlässigendes Potential dar. Die

dynamische Entwicklung in diesem Bereich erfordert die Analyse und Abschätzung der technischen Konzepte und Werkzeuge um mittel- bis langfristig neue wertschöpfende Geschäftsmodelle zu etablieren.

Wien/Potsdam/Bonn im Mai 2014 Tassilo Pellegrini, Harald Sack und Sören Auer

Inhaltsverzeichnis

Teil III Fallbeispiele

Teil I
Grundlagen

Linked Data in Unternehmen. Methodische Grundlagen und Einsatzszenarien

1

Andreas Blumauer

Zusammenfassung

Der Einsatz von Linked Data Technologien im Enterprise Data Management birgt vielschichtige Vorteile in der Aggregation, Verwaltung und Bewirtschaftung von geschäftsrelevanten Datenquellen. Linked Data basierte Datenmodelle sind weniger abstrakt als XML-Schemata oder relationale Datenbankmodelle und sind deshalb besser geeignet, Informationen für Mensch und Maschine in einem Modell zu verknüpfen. Semantische Wissensgraphen können hierbei modular und inkrementell entwickelt werden und mit den Geschäftsanforderungen flexibel mitwachsen. Linked Data Graphen tragen direkt zur Verbesserung der User-Experience bei und generieren Netzeffekte rund um interoperable Datenbestände. Der Beitrag führt diese Aspekte in weiteren Details aus und diskutiert vier Anwendungsfälle für den praktischen Einsatz von Linked Data im Unternehmen.

1.1 Einleitung

Das Konzept von *Linked Data* ist eine praktische Umsetzung des *semantischen Webs*. Die Grundidee dafür geht zurück auf Sir Tim Berners-Lee. Berners-Lee, Direktor des World Wide Web Konsortiums (W3C), hat bereits in seinem Grundsatzpapier „Information Management: A proposal“ [8] Ende der 1980er Jahre eine Entwicklungsstufe des Webs skizziert, in dem Informationsbausteine und Prozesse mit Hilfe smarter Software-Agenten automatisch verlinkt werden. Seither wurde die Entwicklung eben dieses smarteren Webs unter der Schirmherrschaft des W3C vorangetrieben. Dazu wurden zahlreiche Spezifikationen und Standards entwickelt und veröffentlicht, die nun die Grundlage für

A. Blumauer ✉
Semantic Web Company GmbH, Mariahilfer Straße 70, 1070 Wien, Österreich
e-mail: a.blumauer@semantic-web.at

T. Pellegrini, H. Sack, S. Auer (Hrsg.), *Linked Enterprise Data*, X.media.press,
DOI 10.1007/978-3-642-30274-9_1

ein weitreichendes Spektrum an Linked Data Technologien bilden, das von Daten- und Wissensmodellierung über graph-basierte Abfragesprachen bis hin zum automatischen Reasoning reicht.

Neben ihrer technischen Fundierung sind Entwicklungen, vor allem im Umfeld des Internet, dann nachhaltig und zukunftsweisend, wenn sich auf Basis offener Standards auch eine breite Community aus Software-Entwicklern, Beratern und Business-Developern etablieren kann, die die Vorteile ihrer Produkte letztlich auch gegenüber der Industrie demonstriert. Mit dem W3C im Kern wuchs eine solch weltumspannende Community heran, die 10 Jahre nach ihrer Initiierung zu einem Software- und Dienstleistungsmarkt herangereift ist. Das Semantic Web konnte in akademischen Kreisen Fuß fassen und hat sich in unterschiedlichsten Branchen und Industrien als Lösungsmethode für diverse Herausforderungen im Daten-, Informations- und Wissensmanagement etabliert.[1] Semantic Web Technologien ziehen damit in den Alltag ein: sowohl, um Arbeits- und Produktionsprozesse effizienter zu gestalten, als auch bei Entscheidungen oder der Aneignung von Wissen zu unterstützen.

In diesem Überblicksartikel soll zunächst die Idee, auf die sich der Begriff Linked Data bezieht, vermittelt werden. Damit sollen auch LeserInnen angesprochen werden, die sich bislang mit dem semantischen Web bzw. Linked Data nur gelegentlich befasst haben. Daran angeknüpft werden Anwendungsszenarien beschrieben, die den Mehrwert von Linked Data plastisch vor Augen führen. Abschließend wird auf den aktuellen Entwicklungsstand und auf Zukunftsperspektiven von *Linked Enterprise Data* eingegangen.

1.2 Linked Data, Semantic Web, Web of Data – eine Kurzdarstellung

Die Begriffe „Semantische Technologien“ und das „Semantische Web“ werden oft synonym gebraucht, obwohl wesentliche Unterschiede bestehen: Geht es in beiden Fällen darum, Informationen *und* ihre Bedeutung zu verarbeiten, so dienen semantische Technologien der (meist automatischen) Bedeutungserschließung, wohingegen das Semantic Web die bedeutungstragenden Elemente verknüpft und inhaltlich kontextualisiert. Im Semantic Web dreht sich alles um die Frage, wie Entitäten (Produkte, Organisationen, Orte, etc.) sinnvoll zu so genannten *Wissensgraphen* (oder Linked Data Graphen) verwoben werden können. Die zugrundeliegenden Linked Data Technologien setzen dabei auf dem Paradigma der größtmöglichen *Interoperabilität* durch *offene Standards* auf.

Parallel zum allgemein bekannten *Web of Documents* (als Sammlung von HTML-Files), dessen wesentliches Merkmal Hypertext ist, entwickelt sich also ein *Web of Data* (als Sammlung von RDF-Daten), in dem anstelle von Dokumenten Entitäten unterschied-

[1] Über diesen Band hinausgehend dokumentiert insbesondere die englische Literatur zahlreiche Anwendungsbeispiele etwa bei [11], [19] und [20]. Siehe auch die Use Case Sammlung des W3C: http://www.w3.org/2001/sw/sweo/public/UseCases/, aufgerufen am 22.02.2014.

licher Kategorien (z. B. Orte, Organisationen, Produkte, Themen, Personen, . . .), Bezeichnungen, Attribute und deren Relationen zueinander verwaltet werden. Das *Web of Data* wird damit zu einer weltweit verteilten, aber hochgradig vernetzten Datenbank.

Die dem *Web of Data* zugrundeliegenden Design-Prinzipien, wie sie von Tim Berners-Lee [9] definiert wurden, bestehen aus vier Regeln und sind ebenso trivial wie effektiv. Die Prinzipien lauten:

1. Verwende URIs (kurz für: Uniform Resource Identifiers) um Entitäten (Dinge oder Resources) zu bezeichnen.
2. Verwende http-URIs, damit Software-Anwendungen und auch User auf diese Entitäten einfach zugreifen können.
3. Wird eine URI aufgerufen, so liefere nützliche Informationen zurück und verwende dabei offene Standards (RDF, SPARQL).
4. Biete dabei auch Links auf andere URIs an, damit weitere Dinge entdeckt werden können.

Ein kleines Beispiel soll die eben besprochenen (rekursiven) Linked Data Prinzipien veranschaulichen:

1. Tim Berners-Lee hat (unter anderem) die URI http://dbpedia.org/resource/Tim_Berners-Lee
2. Diese URI und damit zahlreiche Fakten zur entsprechenden Entität können von Software-Anwendungen als auch mit einem einfachen Browser aufgerufen werden.
3. Das Resultat ist maschinenlesbar und Standard-basiert. Es werden relevante Fakten in strukturierter Weise retourniert, u. a.:
   - Tim Berners-Lee wurde am 08.06.1955 in London geboren.
   - Tim Berners-Lee heißt auch „TimBL“.
   - Tim Berners-Lee ist der Direktor des „World Wide Web Consortiums (W3C)“.
4. Das W3C hat wiederum eine URI http://dbpedia.org/resource/World_Wide_Web_Consortium, auf die in Folge verlinkt wird, usw.

Abbildung 1.1 zeigt eine Teilansicht des Linked Data Graphen, der sich rund um die URI http://dbpedia.org/resource/Tim_Berners-Lee aufspannt.

Unter Berücksichtigung dieser Linked Data Prinzipien wurde im Jahr 2006 das *DBpedia-Projekt*[2] ins Leben gerufen. Als semantische Version der Wikipedia bildet die DBpedia den Nukleus der stetig wachsenden *Linked Open Data Cloud* (LOD Cloud)[3], eines gigantischen Wissensgraphen, der auch zunehmend kommerziell genutzt wird.

[2] Die öffentlich zugängliche Datenbasis ist mit Stand Februar 2014 in 119 Sprachen verfügbar. Siehe http://dbpedia.org/About, aufgerufen am 22.02.2014.
[3] Siehe http://datahub.io/de/group/lodcloud, aufgerufen am 22.02.2014.

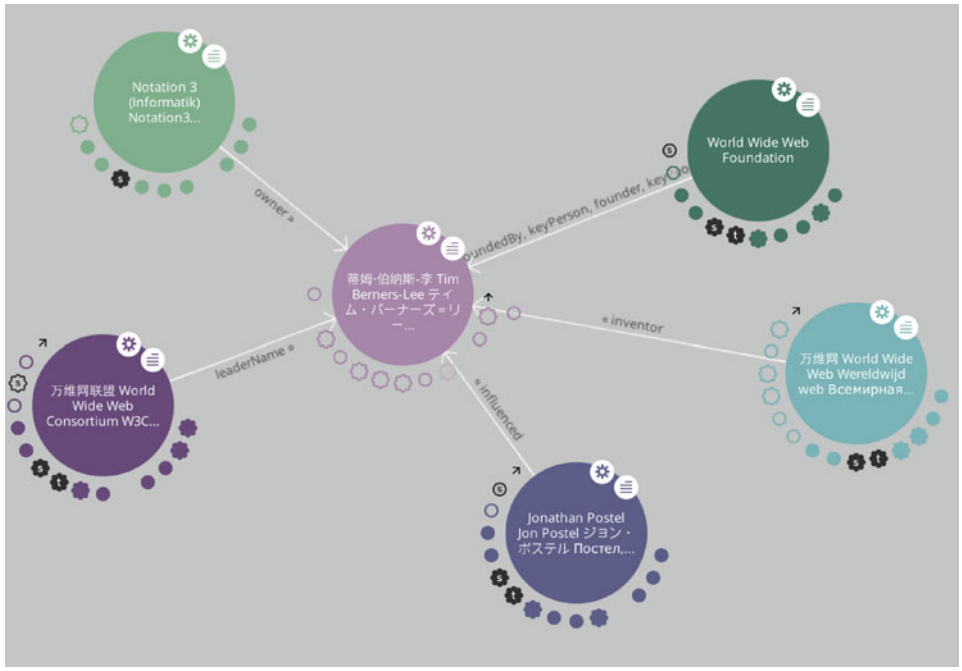


Abb. 1.1 Linked Data Graph zu Tim Berners Lee erstellt mit LodLive (http://www.lodlive.it/), am 19.03.2014

1.3 Linked Data im Enterprise-Einsatz

Unternehmen organisieren Informationen traditioneller Weise mit Hilfe von Informationstechnologien wie relationalen Datenbank-Managementsystemen (RDBMS) oder Dokumenten-Managementsystemen (DMS). Damit werden einerseits strukturierte und andererseits unstrukturierte Informationen gespeichert. Strukturierte Informationen, wie z. B. Kundenstammdaten, folgen einem explizit vorliegenden Datenbank- oder Metadaten-Schema, in dem bereits große Teile der semantischen Information codiert sind. Unstrukturierte Informationen, wie z. B. Gesprächsprotokolle oder Nachrichtentexte, verfügen über keine oder kaum gesondert ausgewiesene Schemata und „funktionieren" auf Basis der Regeln einer natürlichen Sprache wie Deutsch oder Englisch.

Eine wesentliche Idee von Linked Data ist es, dass Daten und Informationen unterschiedlichster Herkunft und Struktur auf Basis von Standards interpretiert, (weiter-) verarbeitet, verknüpft und schließlich dem User in einer Form präsentiert werden können, sodass dieser seine Aufwände zur Informationsgewinnung und -aufbereitung verringern kann.[4] Dementsprechend unterstützen Linked Data Technologien die Datenintegration

[4] Eine umfassende Darstellung des Linked Data Lifecycles findet sich bei Auer et al. [2].

mittels eines ausdrucksstarken Datenmodells, dem so genannten *Resource Description Framework (RDF)*[5], das als Integrationsschicht für die unterschiedlichsten Repräsentationsformen (relationale Datenbanken, XML, natürlich sprachlicher Text etc.) dient. Dazu müssen sowohl Syntax als auch Semantik der zu verknüpfenden Informationen mit Hilfe von Wissensgraphen (u. a. kontrollierte Vokabulare und Ontologien) aufeinander abgestimmt werden.

Gegenüber traditionellen, oftmals auf XML basierenden Techniken zur Datenintegration können zumindest folgende sechs Nutzenargumente angeführt werden, die für den Einsatz von Linked Data sprechen.

1.3.1 Linked Data basierte Datenmodelle sind weniger abstrakt als XML-Schemata oder relationale Datenbankmodelle

Datenbankmodelle werden von Informatikern für Techniker, z. B. für Softwareentwickler formuliert. Menschen verwenden jedoch weder Tabellen, Primärschlüssel noch Normalformen, um ein Modell der Realität zu entwickeln oder sich etwas zu merken. Mit konventionellen Methoden wird Fachwissen – technisch bedingt – oft auf eine Weise repräsentiert, sodass ab einem gewissen Komplexitätsgrad die Nachvollziehbarkeit der semantischen Beziehungen kaum noch gewährleistet ist und spätere Änderungen und Erweiterungen am Modell nur mit hohen Aufwänden möglich sind. Der nachhaltige Nutzen eines Daten- bzw. Wissensmodells, unabhängig vom erfassten Fachbereich, beruht jedoch auch auf seiner Nachvollziehbarkeit und Adaptionsfähigkeit. Hier setzt der graphenbasierte Ansatz des Semantic Web an.

Abstraktes Wissen (oft auch Schema-Wissen genannt) kombiniert mit konkreten Fakten (Faktenwissen) kann in ein semantisches Netz verwoben und auf Basis entsprechender Linked Data Graphen repräsentiert werden. Abstraktes Wissen könnte sich auf einfache Regeln beziehen, z. B. dass Länder stets eine Hauptstadt haben oder Hotels einen Ort. Die Tatsache hingegen, dass der Ort „Wien“ die Hauptstadt von Österreich ist, wird zu den konkreten Fakten gezählt. Beide Arten von Wissen können mittels dem bereits erwähnten Resource Description Framework (kurz: RDF) bzw. mittels RDF-Schema ausgedrückt werden. Es entstehen in Folge unzählige Wissensbausteine (so genannte *RDF-Statements oder RDF-Triples*), die explizit und zunächst unabhängig von jeglicher Anwendung vorliegen können. Ein Triple könnte z. B. ausdrücken, dass eine „Organisation“ einen „Direktor“ hat und, daran in einem weiteren Triple geknüpft, dass „Tim Berners-Lee“ „Direktor von“ „W3C“ ist. Zur besseren Veranschaulichung könnten diese Zusammenhänge wie in Abb. 1.2 visualisiert werden.

Das Prinzip Wissen, Fakten und Modelle von den später darauf zugreifenden Software-Anwendungen zu entkoppeln, führt also dazu, dass diese auch für Nicht-Informatiker verständlicher und einfacher zugänglich werden.

[5] Siehe http://www.w3.org/RDF/, aufgerufen am 22.02.2014.

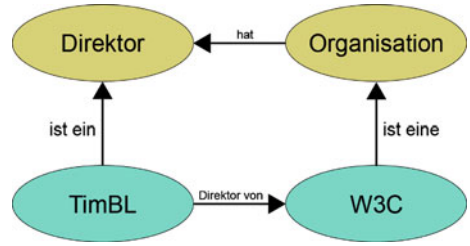


Abb. 1.2 Beispiel für einen einfachen Wissensgraphen bestehend aus vier Triples

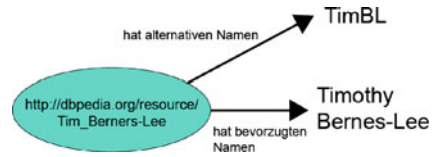


Abb. 1.3 Entitäten, URIs und ihre Labels

1.3.2 Linked Data basierte Datenmodelle verknüpfen Informationen für Mensch und Maschine in einem Modell

Linked Data basiert (im Gegensatz zu relationalen Datenbanken oder XML-basierten Datenstrukturen) auf Graphen, in denen beliebige Entitäten bzw. Ressourcen (z. B. Produkte, Orte, Personen etc.) miteinander semantisch verknüpft werden. Alle Entitäten sind für eine Maschine stets eindeutig über ihren Uniform Resource Identifier (URI) adressierbar. Die abstrakten Zeichenketten der URIs werden für den Menschen mittels „Labels“ (Namensauszeichnungen) lesbar gemacht, wie in Abb. 1.3 veranschaulicht.

Labels können jedoch mehrdeutig sein und sich potentiell auf verschiedene Entitäten beziehen (so genannte „Homonyme“) – ein häufig zitiertes Beispiel hierfür ist „Java“, das sowohl eine Insel, als auch eine Programmiersprache bezeichnen kann. Gibt nun jemand einen Begriff wie „timbl“ in eine herkömmliche Suchmaschine ein, so werden auch Dokumente ausgegeben, die vom „Tilburg Memory Based Learner“ (kurz: TiMBL) handeln. Der Linked Data Ansatz erlaubt durch die Zuordnung eindeutiger URIs die unterschiedlichen Bedeutungen von Labels abzugrenzen. So können Mehrdeutigkeiten bereits während des Indexierens berücksichtigt und leichter abgefangen werden bzw. können Eingabehilfen für den Endanwender bereitgestellt werden.

1.3.3 Semantische Wissensgraphen können modular und inkrementell entwickelt werden und mit den Anforderungen flexibel mitwachsen

Die Rigidität von Schemata in herkömmlichen, relationalen Datenbanksystemen rührt noch von einer Zeit her, als Computersysteme mit einem Bruchteil der heute üblichen Ressourcen und deutlich langsameren CPUs in einer angemessenen Zeit Abfragen über Datenbanken ausführen mussten. Performance wurde „erkauft“, indem das Datenbank-Design zumeist auf wenige spezielle Anwendungsfälle hin optimiert und fixiert wurde.

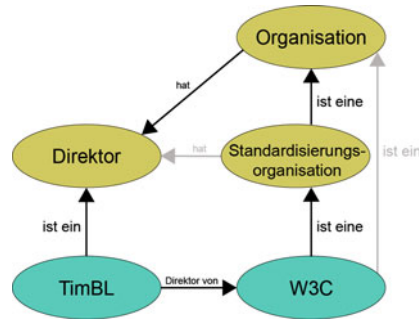


Abb. 1.4 Beispiel dafür, wie ein Wissensgraph Schritt für Schritt erweitert werden kann

Änderungen daran waren hingegen schwierig und wurden tunlichst vermieden, und zwar auch dann, wenn sich die modellierte Realität längst verändert hatte. Diese Starrheit im technischen System hat sich aber nicht nur auf die Geschäftsprozesse, sondern sogar auf das Denken der Software-Entwickler und Datenbank-Manager übertragen: Änderungswünsche an IT-Systemen von Seite der Fachabteilungen, die bis auf die Datenbankebene reichen, werden auch heute noch von IT-Abteilungen gerne als „problematisch“ eingestuft, um es vorsichtig auszudrücken.

Mit dem Aufkommen von leistungsstarken NoSQL- und speziell Graph-Datenbanken, die mit entsprechenden Rechnerleistungen, vor allem aber enormen RAM-Kapazitäten im Server, erst möglich wurden, stellt sich nun allmählich ein Umdenken ein: Daten- und Datenbankmodelle werden als flexibel und verhältnismäßig unaufwändig an die repräsentierte Realität anpassbar gedacht.

Ein Beispiel für die Flexibilität: Als Betreiber eines Informationsportals zu den Themenkreisen „Semantic Web“, „Data & Text Analytics“ und „Big Data“ greifen wir den Wissensgraphen aus Abb. 1.4 auf. Wir fügen eine neue Kategorie von Organisationen ein, nämlich „Standardisierungsorganisation“, um entsprechende Abfragen bzw. Suchfilter zu ermöglichen.

In einem weiteren Triple kann nun die Tatsache hinzugefügt werden, dass das W3C nicht nur einfach eine Organisation ist, sondern eben eine Standardisierungsorganisation.

Dieses Wissensmodell kann also beliebig erweitert werden, sowohl auf Schema- als auch auf Faktenebene, ohne dabei bereits bestehende Software-Anwendungen grundsätzlich zu beeinträchtigen.

Der Wissensgraph aus Abb. 1.2 ist also entsprechend erweitert worden.

Die grau gekennzeichneten Kanten im Graphen und die entsprechenden Fakten (Tripples) können mit Hilfe von automatischem Reasoning und der entsprechenden Ontologie hergeleitet werden:

- Da eine Standardisierungsorganisation eine Organisation ist und jede Organisation einen Direktor hat, hat auch eine Standardisierungsorganisation einen Direktor.
- Da das W3C eine Standardisierungsorganisation ist, ist es auch eine Organisation (im Allgemeinen).

Für zahlreiche Anwendungen, insbesondere in dynamischen Wissensdomänen, benötigen wir Datenmodelle mit einer höheren Flexibilität als jener von relationalen Datenbankmodellen. Linked Data Graphen können ähnlich wie das Strom- oder Straßennetz mit den Anforderungen mitwachsen. Das zugrundeliegende Resource Description Framework (RDF) und RDF-Schema (RDFS), die Web Ontology Language (OWL) bzw. die Abfragesprache SPARQL, jeweils vom W3C standardisiert, bilden dafür die technische Grundlage.[6]

1.3.4 Mit Linked Data können sowohl strukturierte als auch unstrukturierte Informationen semantisch erfasst und verknüpft werden

In vielen Unternehmen wird der Großteil des entscheidungsrelevanten Wissens aus unstrukturierten Informationen gewonnen, z. B. aus E-Mails, Pressemitteilungen oder aus Gesprächsprotokollen. Wenn diese Informationen nun mit möglichst einfachen Mitteln, z. B. mit Fakten aus Produkt-Datenbanken verknüpft werden können, so werden tiefgreifende Analysen zur Wettbewerbssituation oder zu aktuellen Marktentwicklungen möglich.

Linked Data Warehouses, wie z. B. der PoolParty Semantic Integrator[7], können sowohl Daten aus relationalen Datenbanken erfassen, als auch Daten und Fakten aus beliebigen Arten von Texten, um diese beiden Welten schließlich miteinander in Kombination abfragbar zu machen. Abbildung 1.5 veranschaulicht die grundsätzliche Funktionsweise eines solchen Systems.

Eine fundamentale Methode, die dies ermöglicht, beruht auf der automatischen Extraktion von Entitäten bzw. Konzepten aus Texten (Named Entity Recognotion). So könnte das Dokument mit dem Textfragment „Der Erfinder des World Wide Web, Tim Berners-Lee, ist Direktor des W3C“ gemäß dem Wissensgraphen aus Abb. 1.4 analysiert und annotiert werden. Die extrahierten Entitäten sind in vielen Fällen Personen, Organisationen, Produkte oder Orte. In unserem Fall würde der Text mit den beiden Entitäten „Tim Berners-Lee“ und „World Wide Web Consortium“ annotiert bzw. verknüpft werden. Das Dokument kann demzufolge auch automatisch mit den Kategorien „Standardisierungsorganisation“ und „Direktor“ in Verbindung gebracht werden. Diese automatisierbare Umwandlung von unstrukturierten Texten in Linked Data Graphen bildet die Grundlage, um in Folge komplexe Abfragen über Content-Repositorien mit unterschiedlichsten Namen und Bezeichnern bzw. Strukturen (Metadaten- und Kategorisierungssystemen) absetzen zu können.

[6] Für einen Überblick über Semantic Web Standards siehe http://www.w3.org/standards/semanticweb/, aufgerufen am 22.02.2014.

[7] Für einen Überblick über den PoolParty Semantic Integrator siehe http://www.poolparty.biz/portfolio-item/semantic-integrator/, aufgerufen am 22.02.2014.

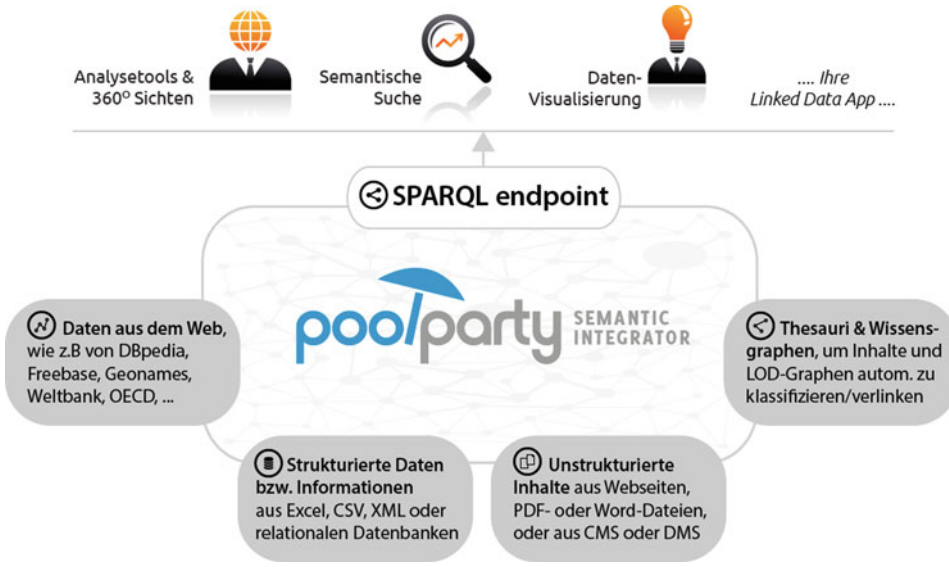


Abb. 1.5 High-Level Architecture eines Linked Data Warehouses, in diesem Beispiel des PoolParty Semantic Integrators

1.3.5 Linked Data besitzt eine mächtige Abfragesprache: SPARQL

SPARQL (kurz für SPARQL Protocol and RDF Query Language)[8] ist ein weiterer wesentlicher W3C-Standard im Semantic Web Stack, mit dem hoch expressive Abfragen von RDF-Graphen ermöglicht werden.

Mit SPARQL können ähnlich komplexe Abfragen wie mit für Business Intelligence Anwendungen typischen OLAP-Würfel ausgeführt werden. Was mit SPARQL nur in wenigen Zeilen formuliert werden kann, würde mit SQL oft nur mit deutlich mehr Aufwand zu bewerkstelligen sein. Die Tatsache, dass mit SPARQL Graphen und nicht Tabellen abgefragt werden, führt dazu, dass z. B. kürzeste Verbindungen zwischen Knoten in einem Netzwerk mit einer einzigen Abfrage ausgeführt werden können.

Seit Veröffentlichung dieses Standards im Jahre 2009 haben sich kaum abweichende SPARQL-Dialekte entwickelt. Dies ermöglicht einerseits die einfachere, herstellerunabhängige Portierung von Daten zwischen RDF-Stores, andererseits können Daten, die sogar in unterschiedlichen Datenbanken gespeichert sind, mit Hilfe von föderierten SPARQL Abfragen „on-the-fly“ und mit einer einzigen Abfrage zusammengefasst und erschlossen werden.

[8] Siehe http://www.w3.org/TR/rdf-sparql-query/, aufgerufen am 24.02.2014.

1.3.6 Linked Data Graphen können direkt zur Verbesserung der User-Experience beitragen

Anwendungen können oftmals nur dann reibungslos bedient werden, wenn dem User das Daten- bzw. Wissensmodell geläufig ist, das der Applikation zugrunde liegt. Mit Hilfe von Linked Data Technologien werden Wissensmodelle leichter zugänglich gemacht. Aus technischer Sicht genügt die Bereitstellung eines SPARQL-Endpoints, um Bedienungselemente wie z. B. Auto-Complete und Auswahloptionen wie Suchfacetten gemäß eines Anwendungs- und User-Kontexts anzubieten.

Wer von uns kennt nicht die prekäre Situation, in der der gewünschte Zug in wenigen Minuten abfährt und der Fahrkartenautomat nur dann das passende Ticket auswerfen würde, wäre man an das Vorhaben mit größerem Vorwissen herangegangen? Ein semantischer Wissensgraph könnte hier – so wie bei jeglichen Suchanwendungen – die erforderlichen Hilfestellungen bereitstellen.

Google's Knowledge Graph (vgl. Abb. 1.6) ist wohl eines der aktuell am sichtbarsten Beispiele, die anschaulich vor Augen führen, wie Linked Data Technologien die User-Experience einer Suchanwendung steigern können. Dabei wird versucht, Suchanfragen in Entitäten, also Ressourcen im Sinne von RDF zu übersetzen. Ist dies erst einmal gelungen, so werden zur initialen Suchanfrage weiterführende Fakten und daran geknüpfte, vertiefende Suchanfragen in einer „Factbox" zusammengefasst. Welche Arten von Fakten angezeigt werden, ist dabei abhängig vom Typ der Ressource. Handelt es sich z. B. um eine Musikgruppe, so werden populäre Musik-Alben der Künstler als Kontextinformation mit ausgegeben, sucht man z. B. nach einer Person wie Tim Berners-Lee, so werden manche seiner Bücher oder seine Eltern als weiterführender Knoten im Wissensnetz zum Anklicken angeboten.

1.3.7 Netzwerkeffekte

Da Linked Data auf einer Vielzahl von offenen Standards beruht, die von Modellierungssprachen wie RDF-Schema und OWL[9], über darauf beruhende Vokabulare und Ontologien wie SKOS (Simple Knowledge Organization System)[10] oder GoodRelations[11] bis hin zur Abfragesprache SPARQL reichen, lassen sich von und für Daten-Provider genauso wie für Endnutzer signifikante Netzwerkeffekte erzeugen.

Primäre Netzwerkeffekte lassen sich dadurch erzielen, dass die Kosten der Informationsintegration und -vernetzung durch standard- und graph-basierte Repräsentation von Information sinken, wobei der Wiederverwendungswert der Information signifikant steigt.[12]

[9] Siehe http://www.w3.org/2001/sw/wiki/OWL, aufgerufen am 24.02.2014.
[10] Siehe http://www.w3.org/2004/02/skos/, aufgerufen am 24.02.2014.
[11] Siehe http://www.w3.org/wiki/WebSchemas/GoodRelations, aufgerufen am 24.02.2014.
[12] Beschreibungen beider Aspekte finden sich bei Cranford [12] und Mitchell & Wilson [16]. Aus Perspektive der Enterprise Search siehe Benghozi & Chamaret [7].

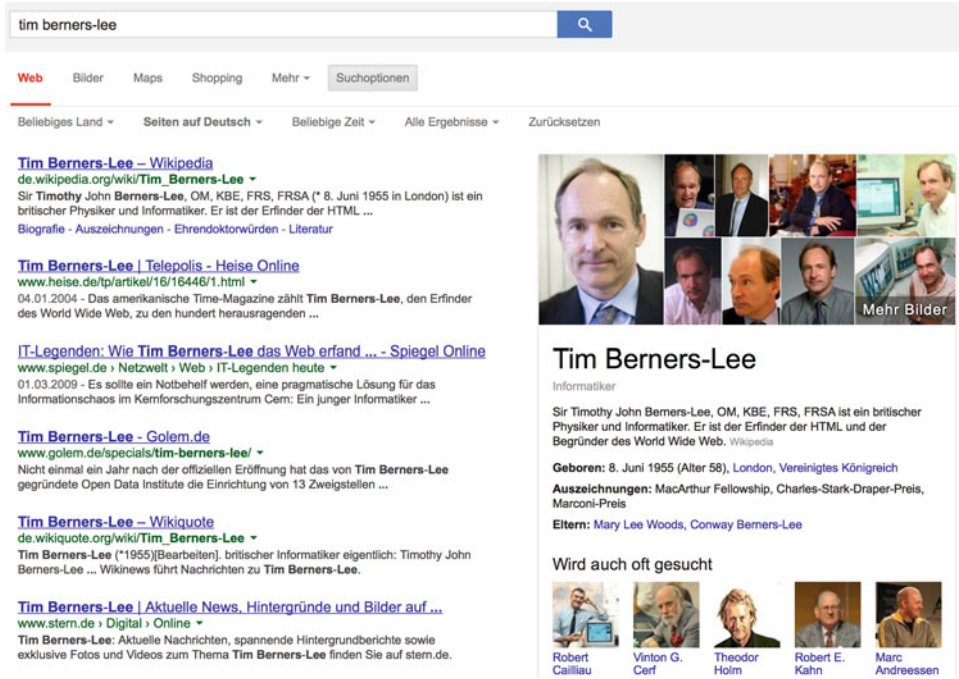


Abb. 1.6 Suche auf Google – die Factbox, rechts neben den herkömmlichen Suchresultaten, wird von Google's Knowledge Graph abgeleitet

Eine anschauliche Entwicklung, die auf diesen Umstand zurückgeführt werden kann, ist das stete Anwachsen der so genannten „Linked Open Data Cloud". Die LOD Cloud[13], die sich zunächst aus zahlreichen, herausragenden Datenbanken und Informationsdiensten wie Wikipedia (bzw. ihrer „semantischen Schwester" DBpedia)[14], Geonames.org[15] oder dem CIA Factbook[16] konstituiert hat, konnte die branchenübergreifende Produktion von Linked Data in unterschiedlichsten Organisationen stimulieren. Prominente Beispiele von Linked Open Data Anbietern sind öffentliche Einrichtungen wie die Europäische Union, die Britische Regierung, Bibliotheken wie die Deutsche Nationalbibliothek oder die Library of Congress, sowie große Medienhäuser wie Wolters Kluwer, New York Times oder BBC.

Aus Sicht von Unternehmen, die auf diesen Zug aufspringen wollen, heißt dies, dass mit jedem Triple, das zur LOD Cloud hinzugefügt wird, der potentielle Wert einer eigenen Linked Data Infrastruktur zunimmt. Dies heißt aber nicht notgedrungen, dass Unterneh-

[13] Siehe http://datahub.io/de/group/lodcloud, aufgerufen am 24.02.2014.
[14] Siehe http://dbpedia.org/About, aufgerufen am 24.02.2014.
[15] Siehe http://www.geonames.org/, aufgerufen am 24.02.2014.
[16] Siehe https://www.cia.gov/library/publications/the-world-factbook/, aufgerufen am 24.02.2014.

men ihre Daten als Linked *Open* Data veröffentlichen müssen. Es können auch interne Linked Data Warehouses um Daten aus der LOD Cloud mit vergleichsweise geringen Aufwänden angereichert werden ohne eigene Daten offenlegen zu müssen.

1.4 Anwendungsszenarien von Linked Data in Unternehmen

Es können zumindest drei grundlegende Szenarien für den unternehmerischen Einsatz von Linked Data unterschieden werden:

1. **Linked Data als Datenintegrationsprinzip anwenden**
   Das Unternehmen verwendet die Linked Data Prinzipien und Semantic Web Technologien intern, um Datenintegration und Mashups (z. B. für ein Wissensportal) zu realisieren bzw. neue Möglichkeiten von semantischer Suche zu erschließen. Dies ist grundsätzlich in allen Geschäftsprozessen bzw. Fachabteilungen von Interesse, in denen durch integrierte Sichten über umfassende, oftmals heterogene und verteilte Datenbestände fundiertere Entscheidungen bei kürzeren Recherchezeiten ermöglicht werden.
2. **Daten aus der Linked Data Cloud einbinden**
   Das Unternehmen konsumiert Daten aus der Linked Data Cloud, um damit z. B. interne Datenbanken oder Inhalte anzureichern. Ein einfaches Beispiel dazu: Werden in den Helpdesk eingehende E-Mails um Geodaten (z. B. von Geonames.org) angereichert, so kann eine Kartenvisualisierung dynamisch erzeugt werden, die anzeigt, aus welchen Regionen zu einem Zeitpunkt die häufigsten Störmeldungen gemeldet werden. Dies kann z. B. für Mobilfunkbetreiber bzw. Stromversorger von Interesse sein.
3. **Daten in die Linked Data Cloud publizieren**
   Das Unternehmen publiziert eigene Daten und Inhalte in die Linked Data Cloud und erschließt sich damit neue Distributionswege und Verwertungskanäle für digitale Assets. Zu diesen Assets gehören in Folge neben Inhalten und Instanzdaten auch die Metadaten, Vokabulare und Wissensmodelle, mit denen diese organisiert werden. Eine klug versionierte Veröffentlichung und Teilverwertung der Metadata-Assets unter kombinierter Verwendung offener und geschlossener Lizenzmodelle ist vor allem für Medienunternehmen oder medienähnlich agierende Unternehmen ein neu aufkeimendes Betätigungsfeld.

Die in Folge detailliert dargestellten Anwendungsfälle orientieren sich an den eben vorgestellten drei Szenarien.

Anwendungsfall 1: Enterprise Search basierend auf Linked Data Enterprise Search funktioniert grundlegend anders als die Suche über Internet-Inhalte. Anders als im WWW kann die Relevanz eines Suchergebnisses nicht auf Basis des Verlinkungs-Grades eines Dokuments berechnet werden, da Firmen-Intranets bei weitem weniger Link-Strukturen

aufweisen als das Internet. Umso wichtiger wird daher die semantische Analyse jedes zu indizierenden Dokuments mit Hilfe linguistischer Verfahren und mit Verfahren des Text-Minings. Damit kann das System nicht nur besser „verstehen“, welche Inhalte ein Dokument hat, sondern auch lernen, wie Begriffe, Phrasen bzw. Entitäten (Orte, Personen, Produkte, Branchen etc.) eines Unternehmens zueinander in Beziehung stehen. Somit können Suchmaschinen z. B. Personen als Experten für gewisse Produkte oder Branchen identifizieren und zusätzlich zu relevanten Dokumenten ausgeben.

Aufbauend auf den eben aufgezählten Grundeigenschaften liegt der Kern eines leistungsstarken Such-Systems für das unternehmerische Umfeld also in der Möglichkeit, Texte und ihre Bedeutung analysieren zu können und mit Hilfe intelligenter Suchdialoge bzw. -assistenten durchsuchbar zu machen. Ein wesentliches Element dabei ist das Erkennen und Extrahieren jener Entitäten (Geschäftsobjekte), die für ein Unternehmen von besonderer Bedeutung sind. Dazu zählen zumeist Produktnamen, Unternehmen (Kunden, Partner, Tochter- und Schwesterfirmen), Projekte, Personen usw. Sind diese erst einmal jedem Dokument zuordenbar, können komplexere Suchanfragen abgesetzt werden, die zumeist schon einem Frage-Antwort-System nahe kommen, z. B.: Wer ist Ansprechpartner für ein bestimmtes Produkt? Oder welche Projekte wurden am Standort X in einem gewissen Zeitraum durchgeführt?

Das Informationsbedürfnis eines Mitarbeiters, der in einer wissensintensiven Branche komplexe Aufgaben zu bewältigen hat, entspricht in vielen Fällen einer Beratungssituation. Nicht eine singuläre Suchanfrage wie „Pizza Wien“, was für die Suche im Web typisch ist, sondern eine Abfolge an Fragestellungen ist zu unterstützen. Diese „moderierte Suche“, die mit Hilfe von Such-Assistenten wie Facetten-Suche und dem so genannten „Drill Down“ ermöglicht wird, gehört zu den aktuellen Features einer Suchmaschine, die auf dem Stand der Zeit ist.

Suchmaschinen der neuesten Generation können nicht nur einzelne Entitäten erkennen und extrahieren, sondern sogar Fakten und Aussagen, und diese können zueinander in Beziehung gesetzt werden. Zum Beispiel wird erkannt, dass Person X in einem gewissen Zeitraum der CEO eines Unternehmens Y war, und weiters, dass dieses Unternehmen Y in den Jahren zuvor das Tochterunternehmen einer Firma Z war usw. Dokumente werden also mittels der extrahierten Entitäten und Fakten mit dem Linked Data Graphen verknüpft, wodurch auch unstrukturierte Informationen reichhaltig kontextualisiert und via der Abfragesprache SPARQL zugänglich gemacht werden können.

Anwendungsfall 2: Mitarbeiterportal Mitarbeiterportale sind wesentlicher Bestandteil eines Wissensmanagement-Systems und bieten für jeden Mitarbeiter vor allem bei der Informationsbeschaffung einen zentralen Anlaufpunkt. Ob nun eine datenbank- und anwendungsübergreifende integrierte Sicht auf die betriebliche Informationslandschaft am Portal erzeugt werden kann, hängt davon ab, ob Doppelgleisigkeiten beim Aufbau von Referenz- und Identifikations-Systemen vermieden werden können.

Ein Beispiel dazu: Wird in der einen Datenbank von „Kunde“ gesprochen, in der anderen aber vom „Klienten“, so beziehen sich zwar beide Bezeichner auf dasselbe Geschäftsobjekt, jedoch bleibt der Maschine diese Beziehung verborgen. Eine übergreifende Suche nach allen Kunden oder die ganzheitliche Sicht auf einen Kunden ist damit nicht möglich. Ausweg aus dieser in der Praxis häufig anzutreffenden Situation kann wiederum ein URI-System bieten: Jedes Geschäftsobjekt ist via Uniform Resource Identifier (URI) eindeutig gekennzeichnet und adressierbar.

Damit ist die Basis zur Entwicklung kontextsensitiver, „mitdenkender“ Widgets für ein Mitarbeiterportal gelegt: Inhalte, die von Mitarbeitern eingestellt werden und über ein Tagging-System annotiert werden, das auf Basis eines SKOS-basierten Thesaurus funktioniert, können mit anderen Inhalten aus dem Intranet intelligent verknüpft werden. So kann z. B. die Suche nach ähnlichen Inhalten realisiert werden, was auch dabei helfen kann, das Rad nicht stetig neu zu erfinden, Doppelarbeiten zu vermeiden und weiterführende Quellen zu erschließen.

Anwendungsfall 3: Content Augmentation Content Augmentation bezeichnet jenen Vorgang, in dem Inhalte, die von Autoren oder Mitarbeitern z. B. im Rahmen eines Enterprise-Content-Management-Systems erstellt werden, mit anderen Inhalten angereichert werden. So können mittels Geo-Daten übersichtliche Kartendarstellungen eingebunden und mit weiterführenden sinnvollen Kontextinformationen kombiniert werden. Die Zusatzinhalte stammen aus Internetquellen wie Wikipedia, aus Nachrichtendiensten, Fachdatenbanken oder aus statistischen Zeitreihen sensorgesteuerter Echtzeitdaten – womöglich in Form von Open Data.

Dies kann einerseits für den User bedeuten, dass dieser gewinnbringende Zusatzinformationen ohne weiteren Rechercheaufwand beziehen kann, andererseits können diese zusätzlichen Daten dazu dienen, die Inhalte mit weiteren Metadaten aufzuwerten, was wiederum zu einer effizienteren Inhaltserschließung führen kann und insbesondere im Kontext von Big Data Anwendungen erfolgskritisch ist.[17]

Anwendungsfall 4: Market Intelligence Mit Hilfe integrierter Sichten und mittels Content Augmentation, der zielgerichteten Anreicherung von Dossiers mit Inhalten aus dem Web oder aus anderen Datenquellen, können u. a. folgende Market-Intelligence-Funktionen unterstützt werden:

1. *Prognosefunktion und Trend Scouting*
   Chancen und Entwicklungen werden durch gezieltes Web-Mining frühzeitig aufgedeckt und antizipiert. Veränderungen des marktrelevanten Umfelds können besser abgeschätzt und deren Auswirkungen auf das eigene Geschäft durch semantisches Trend Mining aufgezeigt werden.

[17] Siehe dazu etwa den McKinsey Report zu Big Data [15].

2. *Unsicherheitsreduktionsfunktion durch verbesserte Kontextualisierung*
   Durch die Präzisierung und Objektivierung von Sachverhalten bei der Entscheidungsfindung wird eine typischerweise schlecht strukturierte Problemstellung besser beherrschbar.
3. *Selektionsfunktion*
   Relevante Informationen können aus der Flut umweltbedingter Informationen besser ausgewählt werden.

1.5 Zusammenfassung und Ausblick

Als kennzeichnende Elemente einer anhaltenden Entwicklung, in welcher das traditionelle Web of Documents mit einem Web of Data verknüpft wird, sind zusammenfassend zu nennen:

1. Die Linked Data Initiative des W3C, das ein einfaches Framework, bestehend aus vier Regeln entwickelt hat, um eine weltweite, verteilte Datenbank, das „Web of Data“ zu realisieren [9].
2. Die Übersetzung der Wikipedia in maschinenlesbares Semantic Web Format unter Berücksichtigung der Linked Data-Prinzipen als Nukleus für ein „Web of Data“ [3]. Die daraus resultierende DBpedia ist in der Zwischenzeit in 119 Sprachen verfügbar und bildet den Nukleus der stetig wachsenden „Linked Open Data Cloud“ (LOD Cloud).
3. Die Verwendung von Uniform Resource Identifier (URIs) aus der LOD Cloud, um in Kombination mit automatischen Text-Extraktionsverfahren Web-Dokumente um Metadaten anzureichern, die im Sinne des Semantic Web quellenübergreifend referenzierbar sind. Dieses Grundprinzip macht auch Google für sich nutzbar, indem auf Basis des Google Knowledge Graphs Webinhalte indiziert und verknüpft werden. Damit können beliebige unstrukturierte Informationen als semantischer Graph repräsentiert werden. Werden URIs aus der LOD Cloud verwendet, also z. B. von DBpedia.org, so werden Inhalte aus dem WWW und in weiterer Folge auch aus dem Corporate Web besser verlinkbar und vergleichbar.
4. Das ursprüngliche „Henne-Ei-Problem“, ohne Semantic Web Daten gibt es keine entsprechenden Anwendungen und so fort, wird nicht nur durch Wrapper und Extraktions-Frameworks [14] überwunden, sondern auch durch die zunehmende Verbreitung der Semantic Web Standards über gebräuchliche Plattform- und Datenbank-Technologien wie Drupal, Wordpress oder MarkLogic, die Metadatenformate wie RDF immer stärker ins Zentrum ihrer Architektur rücken. Parallel dazu propagieren auch Suchdienste wie Google zunehmend die Verwendung von Linked Data Standards wie RDFa oder JSON-LD [18].
5. Die BBC als ein europäisches Leitunternehmen hat 2008 schließlich mit „BBC Music beta“ ein Linked Data Projekt vorgestellt, das aufzeigt, welche neuartigen Verwertungsstrategien für Medienunternehmen mit Hilfe des Semantic Web möglich wer-

den [13]. Die Plattform reichert eigene Informationen um Ressourcen aus MusicBrainz und der Music Ontology an und kann damit nicht nur Mashups mit Wikipedia oder MySpace automatisch generieren, sondern bietet so auch neue kostengünstige Möglichkeiten für andere Plattformen an, um Inhalte der BBC einzubinden. Die BBC fühlte sich nach dem Erfolg dieses Projektes veranlasst, den Einsatz von Linked Data Technologien auszuweiten und setzte u. a. damit das Informationsportal der Olympischen Sommerspiele 2012 in London um [5].

6. Dem BBC-Beispiel folgten zahlreiche weitere Unternehmen, u. a. andere Konzerne aus der Medien- und Verlagsbranche wie Wolters Kluwer oder Elsevier, aber auch Betriebe aus anderen Branchen wie der Automobilindustrie, der Pharmaindustrie oder der öffentlichen Verwaltung [4]. Insbesondere öffentliche Einrichtungen wie Ordnance Survey (UK), die Europäische Union, die Weltbank oder Bibliotheken wie die Deutsche Nationalbibliothek tragen immer mehr zur Verbreitung von Daten auf Basis von Linked Data Standards bei. Das Semantic Web hat also begonnen, Einzug in diverse Branchen und Industrien zu halten.

Obwohl sich das neuartige Gebiet *Linked Enterprise Data* zunächst auf semantische Lösungen für die Probleme in den kontrollierten IT-Umgebungen der Unternehmen konzentriert, ist dies zugleich auch eine wichtige Grundlage, um mit der Zeit den Fokus zu erweitern und Entwicklungen hervorzubringen, die über Unternehmensgrenzen hinweg vernetzt auf globale Dimensionen eines Public Semantic Web skalieren. Eine Schlüsselrolle in diese Richtung nimmt dabei der Linked Data Ansatz ein, der es erlaubt, das Unternehmenswissen mit externen Daten (Linked Open Data Clouds) anzureichern bzw. als „Linked Open Data“ anderen zur Verfügung zu stellen.

Eine vollständig integrierte Sichtweise auf ein Corporate Semantic Web kann dann gelingen, wenn ein Unternehmen als Organisation begriffen wird, die Inhalte, Prozesse und Informationen nicht nur innerhalb der Unternehmensgrenzen produziert und einsetzt, sondern im Sinne eines vernetzten Unternehmens im Ökosystem Internet agiert. Interne und externe Inhalte sinnvoll und kostenschonend zu verknüpfen, kann nur in einem interoperablen Framework wie dem Semantic Web gelingen. Die umfangreiche Nutzbarmachung von Linked Data Technologien für den kommerziellen und industriellen Einsatz ist damit weniger eine Frage der Technologie als vielmehr ihrer organisationalen Verankerung. Entsprechend werden folgende Fragestellungen an Bedeutung gewinnen:

- *Geschäftsmodelle im Semantic Web*
  Wie können Wertschöpfungsmodelle entwickelt werden, die sich vom Rohdaten-Lieferanten bis hin zum Endkunden erstrecken und über geeignete Daten-Transformationen hin zu Linked Data darauf aufbauende Mehrwert-Dienste und Mashups ermöglichen [6]? Wie können im Zusammenhang damit entsprechende Lizenz- und Preismodelle entwickelt werden [17]?