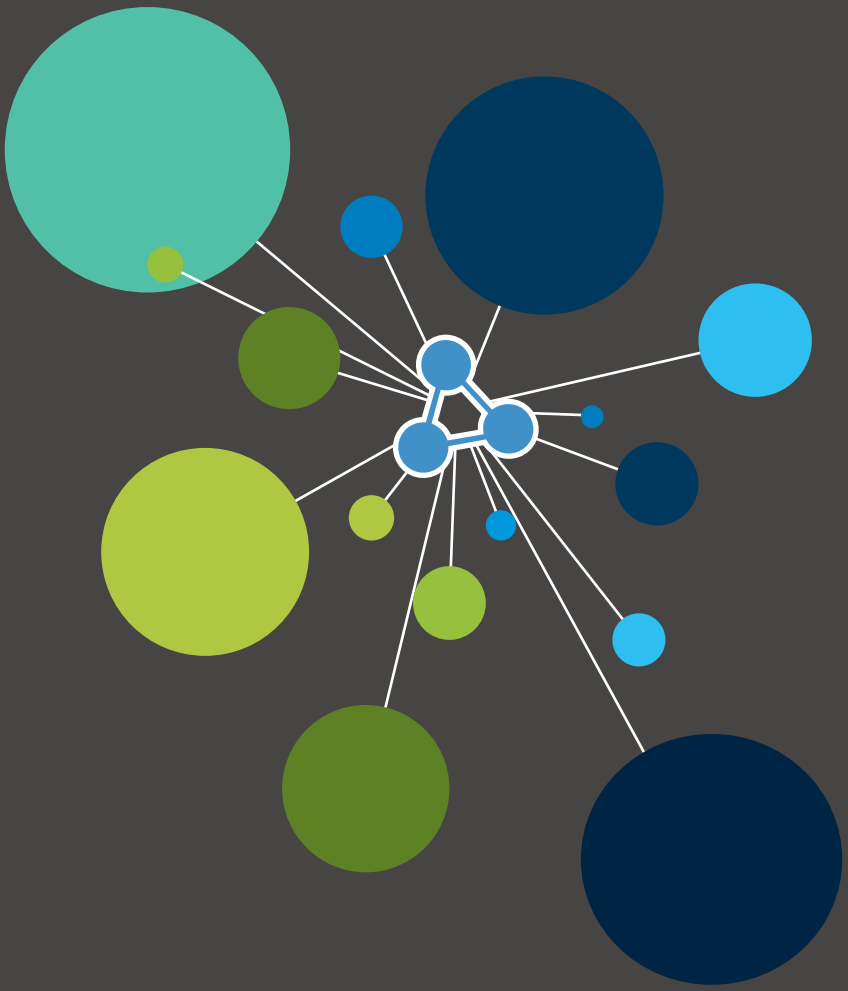
N LDE
NETWORK OF LINKED DATA EXCELLENCE



# Vernetzte *Daten* in der *Verlagsbranche*

*Einsatz, Nutzen und Grenzen von Linked Data Technologien im Enterprise Data Management von Verlagen*

Autoren

Tassilo Pellegrini, Clemens Wolf & Thomas Wolkinger

Diese Veröffentlichung wurde erstellt mit Unterstützung von (i. a. R.)

Eva Goldgruber
Johanna Grüblbauer
Katrin Nussmayr



Umschlagbild & Layout Boris Böttger
Druck DMS DATA+MAIL Schinnerl GmbH

Verlag der FH JOANNEUM Gesellschaft mbH
Alte Poststraße 149
A-8020 Graz
www.fh-joanneum.at

ISBN print: 978-3-902103-60-4
ISBN eBook: 978-3-902103-61-1

# Inhaltsverzeichnis

# Vorwort

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. Dies ist insbesondere für Medienunternehmen relevant, welche moderne Datentechnologien nutzen, um ihr Produktportfolio zu erweitern und neue Formen der Wertschöpfung zu erproben.

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 höchstmögliche 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 dieser Infobroschüre 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 in Medienunternehmen. Hierbei spielt vor allem der Einsatz von Semantic Web Technologien, sowohl als Produktions- als auch Distributionsinfrastruktur für umfassende Datensammlungen, eine zentrale Rolle. Denn durch Semantic Web Technologien werden Daten zu Netzgütern und erlauben neue Formen der Datenhaltung und Bewirtschaftung.

Die vorliegende Broschüre richtet sich an eine technisch interessierte Leserschaft, die sich mit den Grundlagen und Anwendungsmöglichkeiten des Linked-Data-Prinzips für das betriebliche Informations- und Wissensmanagement vertraut machen möchte.

Tassilo Pellegrini & Thomas Wolkinger

# Web of Data 101

## Linked Data – Das Wichtigste in Kürze

*Die Herausforderungen von Big Data / Mit Linked Data von der Informationsflut profitieren / Das „semantische Web" als Vision*

Digitale Technologien haben in den vergangenen Jahren zu einer regelrechten Explosion verfügbarer Informationen und Daten in allen gesellschaftlichen Bereichen geführt. Es sind unvorstellbare Datenmengen, die heute jede Sekunde von Unternehmen wie Google oder von großen Medienhäusern gesammelt, analysiert, verarbeitet und veröffentlicht werden, die auf Finanzmärkten oder in wissenschaftlichen Forschungslabors anfallen, die von Konsumentinnen und Konsumenten produziert werden, wenn sie Transaktionen mit ihren Computern, Mobiltelefonen oder Kreditkarten tätigen.

Als **Big Data*** wird dieses vielschichtige Phänomen bezeichnet. Big Data steht dabei sowohl für eine Realität, die im Übrigen auch für kritische Diskussionen sorgt (Stichwort „Datenschutz"), vor allem aber für eine Verheißung: Verspricht doch die verantwortungsvolle Nutzung dieser Daten – von denen viele als **Open Data** auch noch frei verfügbar sind – überraschende Sichtweisen auf unsere Welt sowie gänzlich neue Geschäftsmodelle und innovative, profitable Anwendungen.

### Ordnung ins Chaos bringen

Mit dem rasanten Anwachsen der globalen Datenmengen stehen Konsumentinnen und Konsumenten und Unternehmen auch vor neuen Herausforderungen. Denn je größer der Datenpool ist, desto schwieriger wird es, schnell und zuverlässig genau die Informationen zu finden, die gerade benötigt werden. Nur ein sehr kleiner Teil aller Daten ist über das World Wide Web abrufbar, viele Daten stecken in proprietären Datensilos fest, andere liegen noch dazu nur in unstrukturierten, schwer lesbaren oder miteinander nicht kompatiblen Formaten vor. Das erschwert die Nutzung durch Konsumentinnen und Konsumenten, das erschwert aber auch die unternehmensinterne Verknüpfung und Nutzbarmachung von Daten unterschiedlicher Quellen. Oft weiß die eine Abteilung eines Unternehmens nicht einmal, welche Informationsschätze eine andere bereits erschlossen hat.

Genau hier setzt Linked Data an. Unter **Linked Data** wird ein Set von Technologien und Praktiken verstanden, die den Nutzen eigener sowie „fremder", freier Daten optimieren. Indem sie Ordnung im Informationschaos schaffen, eigene Daten ordentlich „informieren", also mit eindeutiger Struktur, mit Bedeutung sowie mit Links auf externe Daten anreichern, und damit für Konsumenten ebenso wie für Maschinen besser verfügbar, lesbar und verlinkbar machen. Im Unternehmen verbessern Linked-Data-Technologien Prozesse und Workflows, sie machen bestehende Services und Anwendungen komfortabler nutzbar und ermöglichen darüber hinaus eine Vielzahl völlig neuer Services und Applikationen, die sich die Fülle frei verfügbarer, strukturierter Daten im Internet, also von **Linked (Open) Data**, zunutze machen können. Große Verlage und Medienunternehmen wie die British Broadcasting Corporation (BBC),

> *„Linked data is essential to actually connect the semantic web."*
>
> Tim Berners-Lee

* Begriffe werden im Glossar am Ende des Manuals erklärt

der Springer Verlag oder Thomson Reuters arbeiten bereits erfolgreich mit diesen Technologien.

## Wenn Maschinen verstehen lernen

Ein weiterer Vorteil von Linked Data: Die damit verbundenen Technologien werden vom führenden Standardisierungsgremium des WWW, dem World Wide Web Consortium (W3C), das von WWW-Miterfinders Tim Berners-Lee gegründet wurde, entwickelt und mitgetragen. Das Ziel dieser Bestrebungen: ein **Semantic Web** zu erschaffen, ein Web of Data, das sich wie eine zweite Schicht über das WWW legt, dieses besser strukturiert und leichter erschließbar macht. Denn Linked Data macht den Bedeutungsgehalt von Daten, ihre Semantik, für Maschinen lesbar. Maschinen werden dadurch in die Lage versetzt, besser zu „verstehen", was Nutzerinnen und Nutzer wirklich finden wollen.

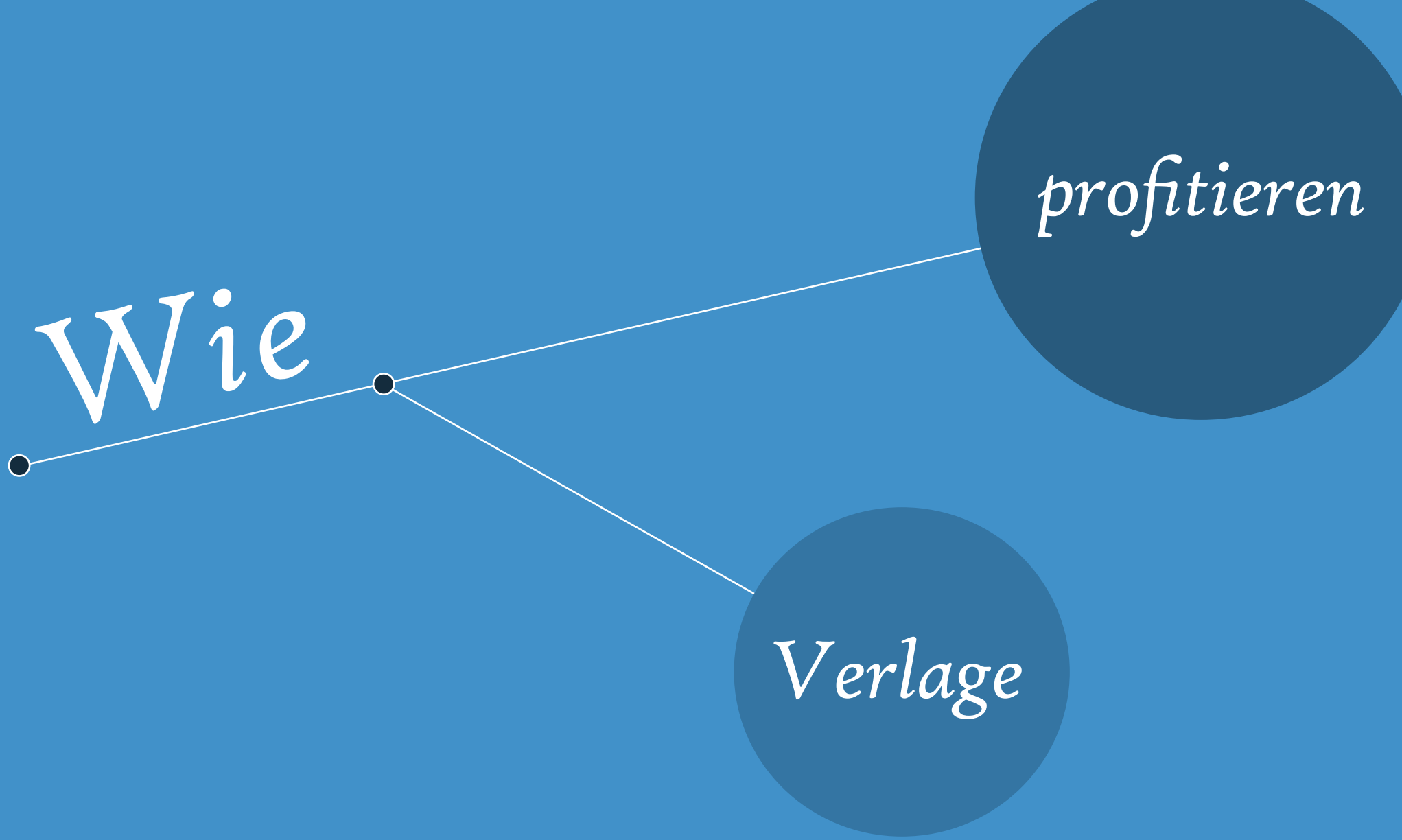
### Lesetipps


Bizer, Christian; Heath, Tom; Berners-Lee, Tim:

Linked Data - The Story So Far. In: International Journal on Semantic Web and Information Systems, 3/2009.
***http://www.ijswis.org/?q=node/31#issue3.***

Mayer-Schönberger, Viktor; Cukier, Kenneth:

Big Data. Die Revolution, die unser Leben verändern wird. München 2013.


Wie
profitieren
Verlage

## So profitieren Verlage von Linked Open Data

*Verlinkt und offen – das neue Daten-Ökosystem /*
*Was es bringt: Die BBC tut es, die New York Times tut es, der Springer Verlag tut es*

Immer mehr Verlage und Medienunternehmen erkennen die Potenziale, die in Linked-Data-Technologien stecken. Kein Wunder, ist doch einerseits der Leidensdruck über die vergangenen Jahre kontinuierlich gestiegen. Immer größere Mengen an unstrukturiertem Content stehen grundsätzlich in den unterschiedlichen Einheiten der Unternehmen zur Verwertung bereit, konnten aber bislang – vor allem mangels Expertise – nicht oder nur unter großem Aufwand weiter verarbeitet und kommerzialisiert werden.

### Hochvernetztes Daten-Ökosystem

Dabei eröffnen zwei Entwicklungen seit Kurzem ganz beachtliche Chancen für die Branche. Da ist zum Einen der seit Jahren stetig wachsende Wissensschatz in Form von **Linked Open Data**, die – im Technik-Jargon ausgedrückt – als **RDF-Daten** strukturiert aufbereitet und untereinander verlinkt vorliegen und bereits vielfach durch Kombination mit unternehmenseigenen Daten zu neuen vermarktbaren Wissenssammlungen und innovativen Anwendungen verbunden werden. Diese **Linked Open Data Cloud** umfasst bereits mehrere Milliarden Fakten aus unterschiedlichsten Themenfeldern – von geographischen Informationen über bibliographische oder statistische Daten bis hin zu spezialisierten Fach-Datenbanken für Musik oder Medizin. DBpedia zum Beispiel, einer der wichtigsten Knoten in diesem hochvernetzten Linked-Data-Ökosystem, stellt RDF-Auszüge aus Wikipedia bereit, mit denen sich bestehender Content sinnvoll anreichern lässt. All diese Daten, die von Unternehmen und öffentlichen Einrichtungen, aber auch von Nutzercommunitys publiziert werden, sind vorwiegend offen lizenziert und werden bereits aktiv kommerziell genutzt.

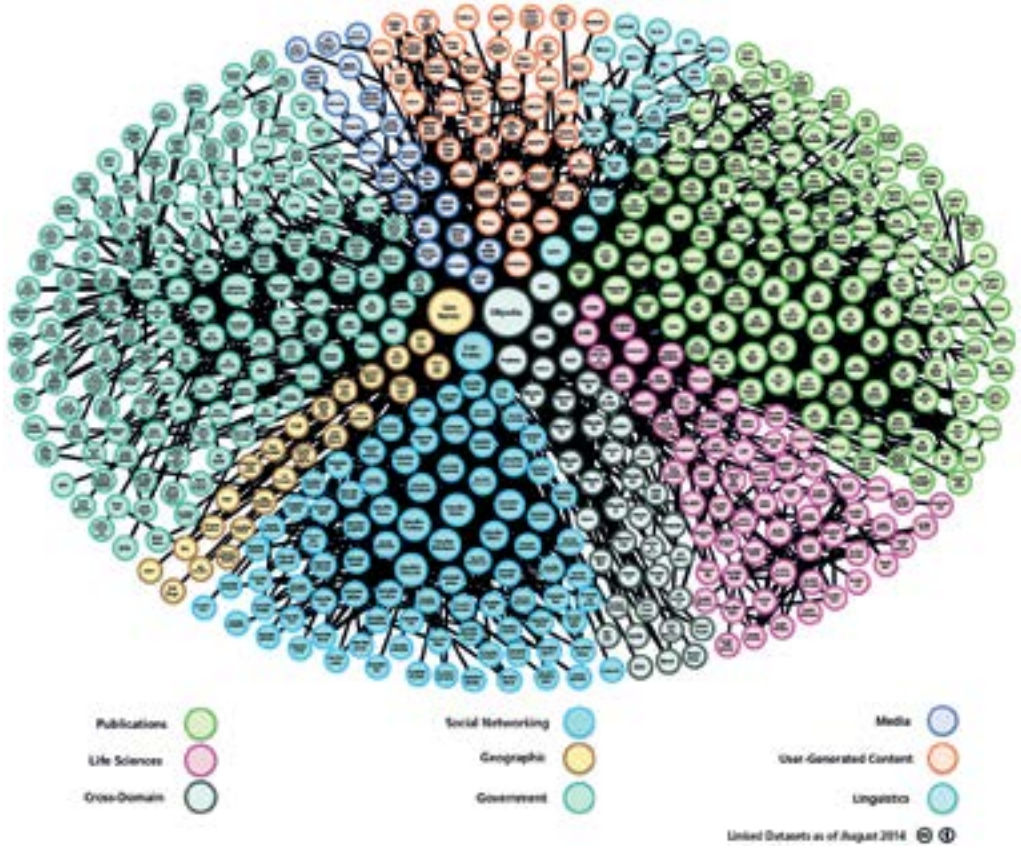



*Abbildung 1: Visualisierung der LOD-Cloud, Stand 30.8.2014,*
*Quelle: http://lod-cloud.net*

Die zweite Entwicklung betrifft Technologien, die es erleichtern, diese Daten auch effizient kommerziell zu verwerten. So wurden seit etwa 2010 sowohl in der automatisierten Verarbeitung von Content (Content Curation) als auch in der automatischen Syndizierung und Integration von strukturierten Daten (Dynamic Semantic Publishing) große Fortschritte erzielt. Diese Linked-Data-Technologien halten sukzessive Einzug in die Redaktionssysteme und Verwertungsstrategien von Unternehmen, die Content verarbeiten. Gerade Verlage und Medienunternehmen wie die BBC, New York Times, Reuters, Reed Elsevier, der Springer Verlag, Agence France Press oder Google haben sich als Early Adopter von Linked Data Technologien hervorgetan.

## Das Jahr, in dem die BBC Kontakt aufnahm

Als öffentlich-rechtliche Rundfunkanstalt mit weltweit mehr als 25 Fernseh- und zwölf Radiosendern sowie einem umfangreichen Internet-angebot produzieren die Redaktionen der British Broadcasting Corporation (BBC) täglich Tausende Sendungen, Formate und Artikel, die zum Teil in völlig voneinander abgeschotteten Content Management Systemen erstellt, mit unterschiedlichen Sets von Metadaten versehen und archiviert werden. „Die Entwicklung eines CMS wird in der Regel vom Publikum der unterschiedlichen Produkte getrieben", schreibt Oliver Bartlett, der die Linked Data-Plattform der BBC mit aufgebaut hat, in einem Blog-Beitrag. „Das bedeutet, dass es üblicherweise nicht dafür optimiert ist, den Content auch für andere Produkte und Services der BBC oder für das WWW zugänglicher zu machen." Wollte man zum Beispiel über alle Plattformen und Formate der BBC die zwanzig zuletzt publizierten Sendungen oder Texte über Burkina Faso finden, so wäre das, meint Bartlett, überaus schwierig.

Mit Linked Data hat sich das geändert. „Plötzlich können wir Verbindungen zwischen allen möglichen Produkten und Contents herstellen, zuvor ging das nur in aufwendiger Handarbeit", schreibt Bartlett. 2010 anlässlich der Fußball WM in Südafrika hat die BBC erstmal mit semantischen Publishing-Technologien experimentiert, 2012 für die Olympischen Spiele in London bereits einen vielfach ausgezeichneten Webdienst auf Linked-Data-Basis konzipiert, der eigene Daten mit denen des offiziellen olympischen Daten-Feeds zu einem umfassenden Service verknüpfte – mit dynamisch generierten und aktualisierten Einzelseiten für alle Länder, Wettbewerbe und Disziplinen sowie für jede und jeden einzelnen der rund 10.000 Athletinnen und Athleten. Seither hat die BBC ihre Linked-Data-Plattform weiter stark ausgebaut, integriert nun auch Linked-Data von DBpedia oder MusicBrainz und veröffentlicht die zugrundeliegenden **Ontologien**, die Wissensmodelle und -architekturen, die Linked-Data strukturieren, auf einer eigenen Plattform.

## Warum sich die New York Times „öffnet"

Das Thema Datenaufbereitung ist für die New York Times (NYT), die Tageszeitung mit der größten Redaktion der Vereinigten Staaten, keineswegs neu: Bereits im Jahr 1913 veröffentlichte die NYT erstmals den vierteljährlichen „The Times Index", eine Auflistung aller Artikel, Namen und Themen, die in den vergangenen drei Monaten in der Zeitung erwähnt worden waren. Diesen Index gibt es heute immer noch – natürlich nicht mehr als gedruckten Bericht, sondern als Datenbank. 2009 begann die NYT damit, diese Daten als Linked Open Data zu veröffentlichen und stellt heute bereits mehr als 10.000 Einträge zur freien Weiterverwertung unter einer Creative-Commons-Lizenz zur Verfügung. Gesucht werden kann in den semantischen Daten der New York Times nach Menschen,

Organisationen, Orten und Schlagworten. Zusätzlich können alle Einträge als **SKOS-Daten** heruntergeladen werden.

Das Ziel: Mit den „Times Topic Pages" betreibt die NYT ein thematisch gegliedertes, kostenlos zugängliches Archiv, das bis ins Jahr 1981 zurückreicht. Die Kategorisierung der Einträge erfolgt über insgesamt rund 30.000 Schlagworte, die ebenfalls als Linked Open Data zugänglich sind. Nutzerinnen und Nutzer sollen – im Sinne des Open-Innovation-Gedankens – mit den semantischen Daten der NYT eigene Linked-Data-Anwendungen kreieren können: Als Vorbild lieferte die NYT 2010 eine eigene Web-App namens „Who Went Where", die Nutzern anzeigt, in welchem Zusammenhang die Zeitung über Alumni eines bestimmten Colleges bzw. einer bestimmten Universität berichtet hatte. In einer über Google-Gruppen organisierten Community bietet die NYT ihren Usern außerdem ein Forum für Kommentare, Fragen und Vorschläge zu dieser Linked-Open-Date-Initiative. Und im Blog „Open" schreiben Entwickler aus dem Team der NYT regelmäßig über neue Initiativen zu LOD.

## Wie der Springer-Verlag nachhaltig Geld verdienen will

Obwohl es viele Initiativen gibt, die Daten über wissenschaftliche Publikationen im semantischen Web verfügbar machen wollen, existieren die meisten dieser Daten nach wie vor bestenfalls im Freitext-Format. Strukturierte Daten würden aber der Forschung, Verlagen, Bibliotheken und Sponsoren gleichermaßen helfen, meint der Springer-Verlag und hat deshalb 2013 ein Linked-Open-Data-Pilotprojekt gestartet. Konkret geht es um Tagungsberichte von anfangs 850 Informatik-Konferenzen, die als LOD zur Verfügung gestellt werden sollen.

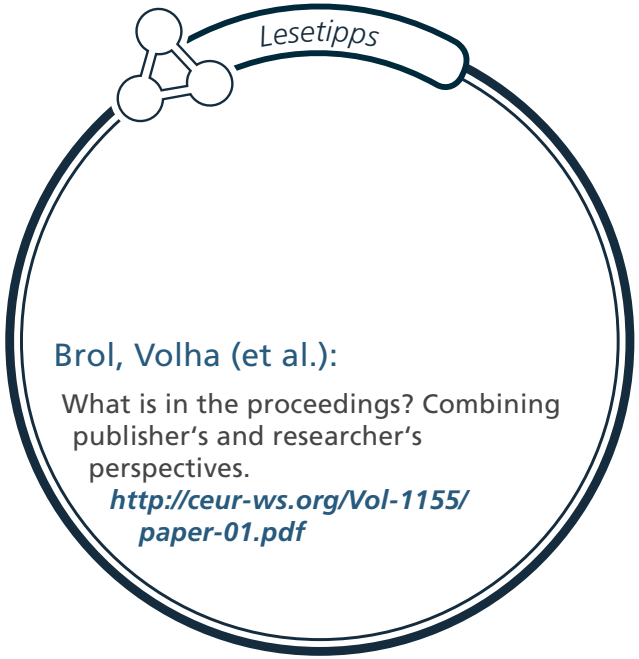
Von seinem Pilotprojekt erwartet sich der Springer-Verlag bessere Zugänglichkeit zu Daten und mehr Transparenz. Und wirtschaftliche Vorteile: Die Verlinkung bringt mehr Leserinnen und Leser, mehr Downloads und damit für die Konferenzveranstalter auch mehr Einreichungen und Teilnehmer. Verlage können neue Einblicke gewinnen und Trends im Konferenz-Geschäft erkennen, auf die dann entsprechend reagiert werden kann. Das Wichtigste sei aber die Nachhaltigkeit, also die längerfristige Nutzbarkeit der Daten, schreiben Vertreter des Springer-Verlags und der Universität Mannheim in einem Paper zum Pilotprojekt des Verlages. Denn sie sei direkt mit dem ökonomischen Wert der Daten verbunden.

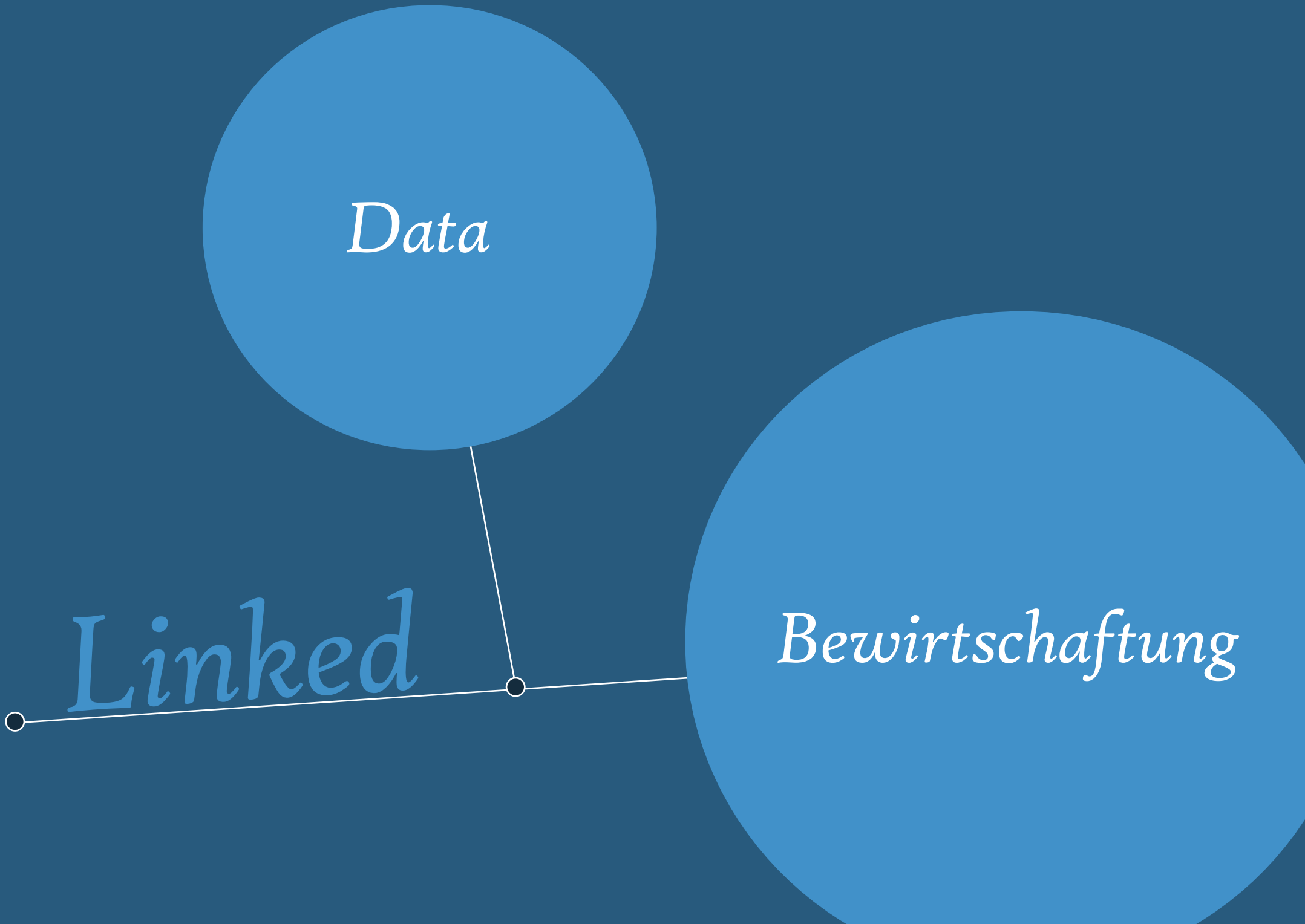
Geholfen ist nach Ansicht von Springer mit einem solchen LOD-Modell allen Beteiligten: Junge Forscher bekommen eine Orientierungshilfe (ob sie ein Paper für eine bestimmte Konferenz einreichen sollen), ältere Forscher können besser selektieren (etwa, ob sie Einladungen zu Konferenzen annehmen sollen) und Verlage können das Potenzial von Konferenzen besser beurteilen (ob es sich auszahlt, die Resultate zu veröffentlichen).

*Linktipps*

**Open**
Datenblog der New York Times
***http://open.blogs.nytimes.com***

**Who Went Were**
Linked-Open-Data-Applikation der New York Times
***http://data.nytimes.com/schools/schools.html***


Lesetipps
Brol, Volha (et al.):
What is in the proceedings? Combining publisher's and researcher's perspectives.
http://ceur-ws.org/Vol-1155/paper-01.pdf


Data
Linked
Bewirtschaftung

## Wie aus Linked Data ein Geschäftsmodell werden kann

*Die neue Content-Wertschöpfungskette / Lizenzierungsmodelle für LOD / Eine Verlagsstrategie entwickeln*

Das Problem unstrukturierter (z.B. MS-Word, PDF) und semistrukturierter Formate (z.B. HTML, CSV), wie auch herstellereigener Datenschnittstellen (APIs) ist bekannt: Mit jeder zusätzlichen Quelle steigt der Aufwand, die Daten zu integrieren. Möchte man die Struktur der Datenbank verändern oder den Bestand aktualisieren, geht das nur unter hohem Aufwand und oft zulasten der Systemintegrität. Genau hier setzt Linked Data an, indem es die Interoperabilität von Datenbanken und IT-Systemen verbessert. Praktisch ermöglicht Linked Data mehr Kosten- und Zeiteffizienz im Datenmanagement und in der Datenbewirtschaftung.

Für den Einsatz von Linked Data in Unternehmen gibt es drei Anwendungsbereiche: 1) Linked Data kann als sogenanntes „Datenintegrationsprinzip" angewendet werden: Die Technologien kommen unternehmensintern zum Einsatz, um Daten zu sammeln und miteinander zu vernetzen, etwa in Form eines Wissensportals oder semantischer Suchfunktionen. 2) Zusätzlich können Unternehmen Daten aus der Linked Data Cloud einbinden und damit ihre eigenen Inhalte anreichern. 3) Weiters können Daten in die Cloud publiziert und damit neue Distributionswege für eigene Inhalte erschlossen werden.

## Linked Data in der Content Value Chain

Dank Linked Data ist es möglich, den digitalen Content entlang der gesamten Wertschöpfungskette kosteneffizient zu bewirtschaften. Von der Content-Akquise, also der Sammlung und Speicherung von Daten, über das Content-Editing und -Bundling, also die Analyse und Kontextualisierung der Daten, bis hin zur Content-Distribution und letztendlich dem Konsum liefert Linked Data technische Möglichkeiten für die entsprechende Datenverarbeitung. Abbildung 2 veranschaulicht den Einsatz von Linked Data entlang der Content Value Chain.

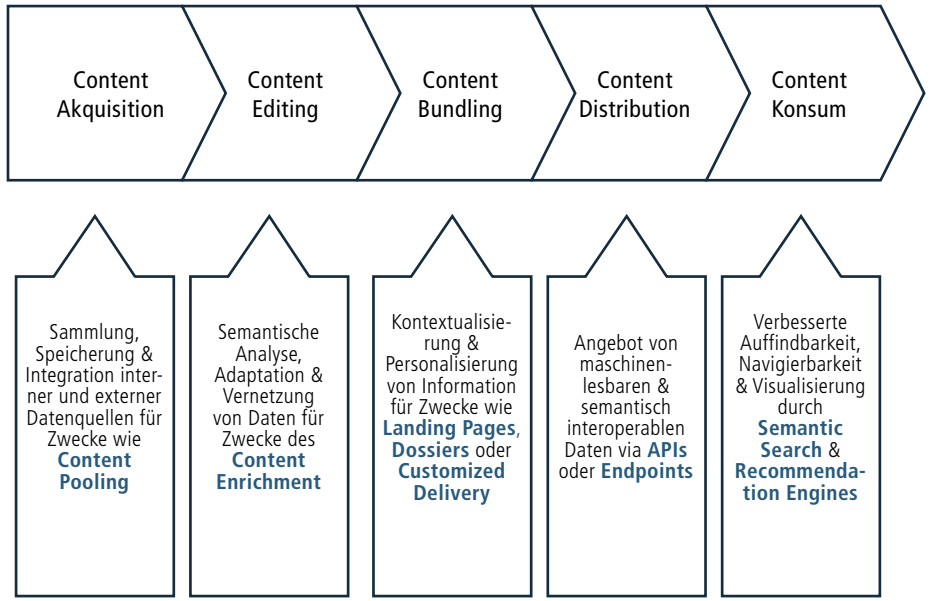



*Abbildung 2: LInked Data in der Content-Value-Chain*

- Die Content-Akquisition umfasst alle Aktivitäten der Sammlung, Speicherung und Integration von Daten. Im Zuge dieses Prozesses werden Fakten und Information von internen und externen Quellen für die weitere Verarbeitung in Form semantischer Indices zusammengetragen.
- Die Stufe des Content-Editings umfasst die semantische Analyse, Adaption, Verlinkung und Qualitätssicherung von Daten für Zwecke des Content Enrichments.
- Das Content Bundling beschäftigt sich mit der Kontextualisierung und Personalisierung von Information. Es dient dem maßgeschneiderten, thematisch kohärenten Zugang zu Content-Einheiten z.B. in Form von Landing Pages oder Dossiers.
- Im Zuge der Content Distribution werden maschinenlesbare und semantisch interoperable Daten z.B. via Programmierschnittstellen oder **SPARQL**-Endpoints für den internen und/oder externen Gebrauch technisch zugänglich gemacht.
- Die Ebene des Content Konsums umfasst alle Maßnahmen der adäquaten Präsentation und Zugänglichmachung von Content in Form von semantischer Suche, Empfehlungs- und Filterdiensten mit dem Ziel, die Auffindbarkeit, Navigierbarkeit und Wiederverwendbarkeit zu erhöhen.

## Der Linked Data Business Cube

Neben der Frage der Lizensierung stellt sich natürlich die grundsätzliche Frage nach dem wirtschaftlichen Wert von Linked Data. Der Linked Data Business Cube ist ein systematischer Ansatz, um geeignete Geschäftsmodelle für Linked Data zu finden und auf unterschiedliche Zielgruppen abzustimmen. Dieses Modell setzt Linked Data Assets – also Informationsbestandteile, die sich als geistiges Eigentum schützen lassen –, mögliche Erlösformen und Stakeholder miteinander in Beziehung. Der Cube kann dabei helfen, verschiedene Linked-Data-Geschäftsmodelle zu unterscheiden und das Leistungsspektrum bzw. den ökonomischen Nutzen eigener Anwendungen besser abzuschätzen.

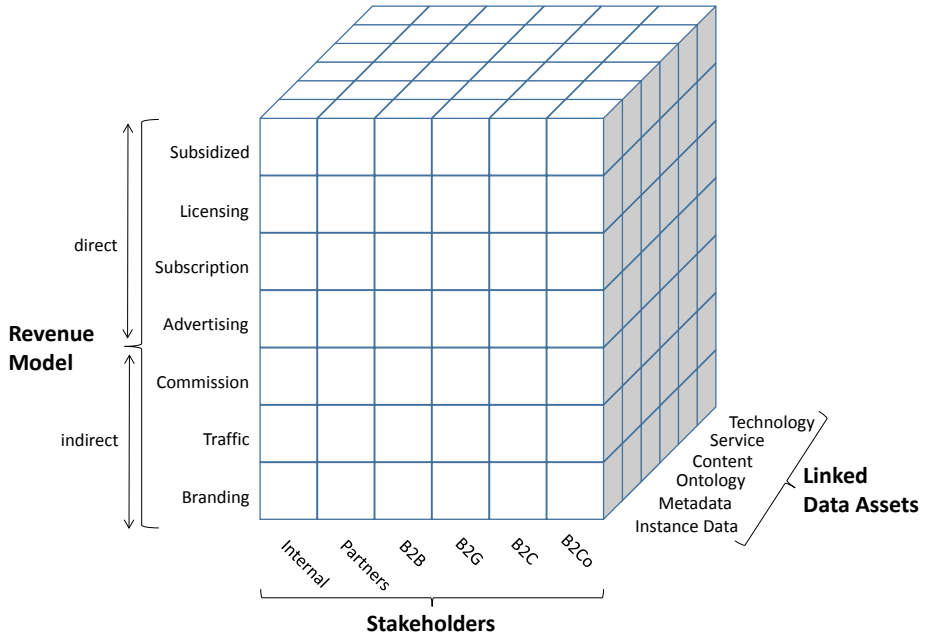



*Abbildung 3: Visualisierung Linked Data Business Cube*

## Praktische Nutzung vernetzter Daten

Das wohl bekannteste konkrete Anwendungsbeispiel des Linked-Data-Paradigmas ist die seit 2007 stetig wachsende „Linked Open Data Cloud", eine kollaborativ gewachsene Umgebung aus Linked Data Quellen, die vorwiegend offen lizenziert sind. Viele Medienunternehmen nutzen diese Daten bereits kommerziell, wie etwa Fallbeispiele der BBC, NYT, Reuters, Springer Verlag, Google oder Facebook dokumentieren.

Doch die Linked Open Data Cloud stellt nur einen Teilbereich des „Web of Data" dar, um das sich bereits ein reger Markt etabliert hat. Diverse öffentliche Stellen veröffentlichen etwa im Rahmen von „Open Government Data"-Initiativen große Mengen an Daten im Netz. Und auch private und stiftungsgetragene Datenanbieter wie datahub.io, Wikipedia, Musicbrainz oder Geonames veröffentlichen ihre Daten zum Teil als Linked Data im World Wide Web. In beiden Fällen haben sich dafür tragfähige Geschäftsmodelle etabliert, die zumeist auf dem Freemium-Modell – also der kostenfreien Nutzung des Basis- und der kostenpflichtigen Nutzung des Vollprodukts – aufbauen. Zusätzlich bieten spezialisierte Aggregatoren wie Socrata, Factual oder QLIK Dienstleistungen rund um deren kommerzielle Nutzung an.

Die Geschäftsgrundlage für die Datenbewirtschaftung im Web of Data bilden Lizenzvereinbarungen zwischen Datenanbieter und Datennutzer. Diese leiten sich aus diversen Schutzrechten des Immaterialgüterrechts ab, die im Zuge der Datengenerierung und -verwaltung geltend gemacht werden können.

## Rechtsschutz von Linked Data

Die Lizensierungsfrage von Linked Data ist nicht trivial, zumal auf die unterschiedlichen Bestandteile eines Linked-Data-Systems unterschiedliche Rechtsgrundlagen anwendbar sind. Im Regelfall sind Datensammlungen in Form von Datenbanken in Europa durch das Urheberrecht gedeckt. Während das Urheberrecht den „kreativen Gehalt" dieser Daten schützt, stellt das europäische Datenbankrecht zusätzlich einen Leistungsschutz dar – und zwar für den Investitionsaufwand, der im Zuge der Sammlung, Verwaltung und Zur-Verfügung-Stellung der Daten anfällt. Daneben kann das Wettbewerbsrecht zur Anwendung kommen, wenn Daten bzw. Bestandteile eines Datensystems missbräuchlich verwendet werden. Diese drei Rechtsbereiche spielen in Folge die wichtigste Rolle für den Entwurf von Licensing Policies für Linked Data und darauf aufbauende Verwertungsmodelle.

Um Netzeffekte und eine möglichst weite Verbreitung zu erzielen, empfiehlt es sich, Linked-Data-Inhalte bzw. deren Bestandteile auch entsprechend differenziert zu lizenzieren. Generell gilt, dass traditionelle, „starke Eigentumsrechte" sich mit dem selbstorganisierenden und dezentralen Charakter des World Wide Web nur eingeschränkt vertragen. Deshalb kommen unter dem Label „Open" vermehrt Commons-basierte bzw. offene Lizenzmodelle zum Einsatz, wobei diese auch mit geschlossenen Lizenzen in Form eines „Dual Licensings" kombiniert werden können.

Im Bereich des Urheberrechts hat sich mit Creative Commons eine tragfähige Alternative für den Schutz von Datenbanken etabliert. Creative Commons kann als Lizenzbaukasten verstanden werden, der eine individuelle und relativ genaue Definition der Nutzungsrechte zulässt – vom Verzicht auf alle Urheberrechte (CC0-Lizenz) über verschiedene Freiheitsgrade für nichtkommerzielle und auch kommerzielle Nutzung. Im Bereich

des Datenbankrechts arbeiten unterschiedliche Initiativen parallel zu Creative Commons an sogenannten Data Commons, einem Set von Lizenzen, das für die Spezifika der Datenlizenzierung optimiert ist.

## Linked Data rechtssicher nutzen

Wiederverwendbarkeit ist eines der Kernprinzipien von Linked Data. Rein technisch gesehen gibt es dabei auch kaum Grenzen: Daten aus verschiedenen Quellen können beliebig miteinander kombiniert werden, um bestehende Anwendungen anzureichern oder aus den erschlossenen Daten neue Produkte bzw. Dienstleistungen zu generieren. Neben der technischen Machbarkeit sind es aber vor allem juristische Fragen, die die Nutzung und Wiederverwendung von Linked Data beeinflussen. Neben dem finanziellen Aufwand für die technische Integration entstehen bei der Datenbewirtschaftung auch Transaktionskosten, und diese können im Falle eines Rechsstreits erheblich ausfallen. Problematisch sind im Zusammenhang mit Rechtefragen vor allem unklare bzw. unvollständige Informationen über die Lizensierung, die den bereitgestellten Daten zugrunde liegen.

Bei der Verwendung fremder Daten gilt es daher zu beachten, inwieweit für diese Lizenzen vorhanden sind. Denn das Nichtvorhandensein einer Lizenz heißt nicht, dass die Daten auch uneingeschränkt verwendet werden dürfen. Im Falle einer Abmahnung kann das zu unangenehmen Folgekosten führen. Was die Bereitstellung eigener Daten anbelangt, empfiehlt es sich, eine vollständige Linked Data Licensing Policy bereitzustellen, die sowohl maschinenlesbare Lizenzen für urheber- und datenbankrechtliche Aspekte enthält als auch eine „Community Norm", die diese Lizenzen für den User verständlich macht und Nutzungsbedingungen im Hinblick auf das Wettbewerbsrecht definiert. Als Repräsentationsstandard für maschinen-lesbare Lizenzen sei auf die **Open Definitions Rights Language** (**ODRL**) verwiesen.

Sowohl bei der Nutzung als auch bei der Bereitstellung von Daten sollte außerdem darauf geachtet werden, dass einzelne Lizenzbestandteile miteinander kompatibel sind bzw. sich nicht gegenseitig ausschließen. Das kann passieren, wenn unterschiedliche Teile eines Datensatzes unterschiedlich lizensiert sind bzw. User selbst Daten aus unterschiedlichen Quellen zu einem neuen Datensatz kombinieren möchten. Wenn die Lizensierung sich nicht eindeutig klären lässt, ist der Datennutzer verpflichtet, mit dem Anbieter Rücksprache zu halten. Es ist in solchen Fällen auch möglich, Nutzungsbedingungen in Form einer individuellen Vereinbarung zu klären. Datensätze, über deren Lizensierung man im Unklaren ist, sollten im Zweifelsfall aber nicht verwendet werden.

*Linktipps: Die ergiebigsten Datenquellen*

**„Creative Commons“**
Hilft Wissen mit der ganzen Welt zu teilen
***http://creativecommons.org/***

**„Open Data Commons“**
Sammlung von Hilfsmittel, um Daten anzubieten und zu nutzen
***http://opendatacommons.org/***

**„ODRL“**
Open Digital Rights Language: Standard um Rechteinformationen von Inhalten zu beschreiben
***http://www.w3.org/TR/odrl/***

**„Linked Data Business Issues“**
Informationen zu Linked Data Business von der Semantic Web Company
***https://blog.semantic-web.at/tag/business-model/***

*Lesetipps*

Pellegrini, T. (2015):

Lizenzierung und Nutzung vernetzter Daten – Fallstricke und Empfehlungen. In B. Ege, B. Humm, & A. Reibold (Hrsg.), Corporate Semantic Web (S. 381–396). Berlin, Heidelberg: Springer Berlin Heidelberg
***http://link.springer.com/10.1007/978-3-642-54886-4_26***

# Relevante Datenquellen aus dem Web: Ein reicher Datenschatz

*Die Linked-Open-Data-Cloud / Von verlinkten Enzyklopädien, Geodaten, Medien und Bibliotheken*

Knapp 90 Milliarden **RDF**-Tripel aus mehr als 3300 Datensätzen: Im Web of Data ist mittlerweile (Stand: Juni 2015) eine beinah unvorstellbar große Anzahl an Daten vorhanden, die nur darauf warten, von Verlagen genutzt zu werden. Von enzyklopädischen Daten über Biomedizin bis hin zu Behördendaten reicht die Bandbreite. Die wichtigsten Datenquellen der LOD-Cloud im Überblick.

## Enzyklopädien

Zu einem wichtigen Knotenpunkt in der LOD-Welt hat sich DBpedia entwickelt: Die Datensammlung bezieht automatisiert Inhalte aus der Online-Enzyklopädie Wikipedia und macht diese in Form von Linked Open Data verfügbar. Herangezogen wird normalerweise die englischsprachige Version von Wikipedia, über sogenannte Interlanguage-Links sind aber auch andere Sprachen eingebunden. Die Daten von DBpedia sind beispielsweise auch mit anderen Diensten wie Geonames, US-Zensusdaten, EuroStat oder dem CIA World Fact Book verbunden.

## Geodaten

Die meistgenutzte Datenquelle für geografische Daten ist Geonames. Der Datensatz beinhaltet über 10 Millionen geografische Namen von Orten auf der ganzen Welt. Geonames referenziert außerdem weitere geografische Datensätze und stellt so auch Informationen über Ortsnamen in anderen Sprachen, Höhe der Orte über dem Meeresspiegel oder Bevölkerungszahl zur Verfügung. Geonames wird von namhaften Unternehmen und Services eingesetzt, darunter vom Betriebssystem Ubuntu, vom Kartendienst Bing Maps, von den Sportartikelherstellern Adidas und Nike sowie von Medien wie der New York Times, BBC oder Norwegian Broadcasting Corp.

## Medien

Im Medienbereich haben sich besonders zwei freie Datensätze durchgesetzt, die Musikdaten referenzieren: BBC Music und MusicBrainz. BBC Music aggregiert musikbezogene Inhalte von allen Websites der BBC und beinhaltet Rezensionen sowie Berichte über Neuveröffentlichungen aus diversen Genres. MusicBrainz sammelt unterschiedlichste Informationen über Musikschaffende und deren Veröffentlichungen. MusicBrainz wird beispielsweise von einer Reihe von Programmen zum Tagging (Kategorisierung) von Musik und virtuellen Medienbibliotheken unterstützt.

## Bibliotheken

Auch im Bereich bibliografischer Daten gibt es mittlerweile einige LOD-Initiativen. Vorreiter war in diesem Bereich der schwedische Verbundkatalog LIBRIS, der seit 2008 als Linked Open Data bereitgestellt wird. Zwei Jahre später stellte die Deutsche Nationalbibliothek ihre Normdaten online. Die Library of Congress publiziert ihre „Subject Headings" und „Control Numbers" als LOD. Und auf internationaler Ebene führt das Virtual International Authority File die Normdaten verschiedener Nationalbibliotheken in einer Datenbank zusammen und stellt diese als Linked

Data zur Verfügung. Als Fachthesaurus für Wirtschaftsliteratur versteht sich der „STW Thesaurus" des Leibniz-Informationszentrums Wirtschaft, der dessen Literaturbestand in Form von LInked Data bereitstellt.

*Linktipps: Die ergiebigsten Datenquellen*

***DBpedia,*** *http://dpbedia.org*

***Geonames,*** *http://geonames.org*

***MusicBrainz,*** *http://musicbrainz.org*

***BBC Music,*** *http://bbc.co.uk/music*

***LIBRIS,*** *http://libris.kb.se*

***Library of Congress Linked Data Service,***
*http://id.loc.gov*

***Linked Data Service der Deutschen Nationalbibliothek,***
*http://www.dnb.de/DE/Service/DigitaleDienste/LinkedData/linkeddata_node.html*

***Virtual International Authority File,*** *https://viaf.org*

***STW Thesaurus,*** *http://zbw.eu/stw/version/latest/about*

*Lesetipps*

Jentzsch, A. (2014):

Linked Open Data Cloud. In T. Pellegrini, H. Sack, & S. Auer (Hrsg.), Linked Enterprise Data (S. 209–219). Berlin, Heidelberg: Springer Berlin Heidelberg.
***http://link.springer.com/10.1007/978-3-642-30274-9_10***

## Interview: Compass-Verlag

*„Spannend wird es, wenn man Daten verknüpft“*

*Interview mit Nikolaus Futter, Geschäftsführer des Compass-Verlags.*

**Welche Rolle spielt Datenbewirtschaftung im Verlagswesen? Markieren die aktuellen Entwicklungen einen Sprung im Vergleich zu früher?**

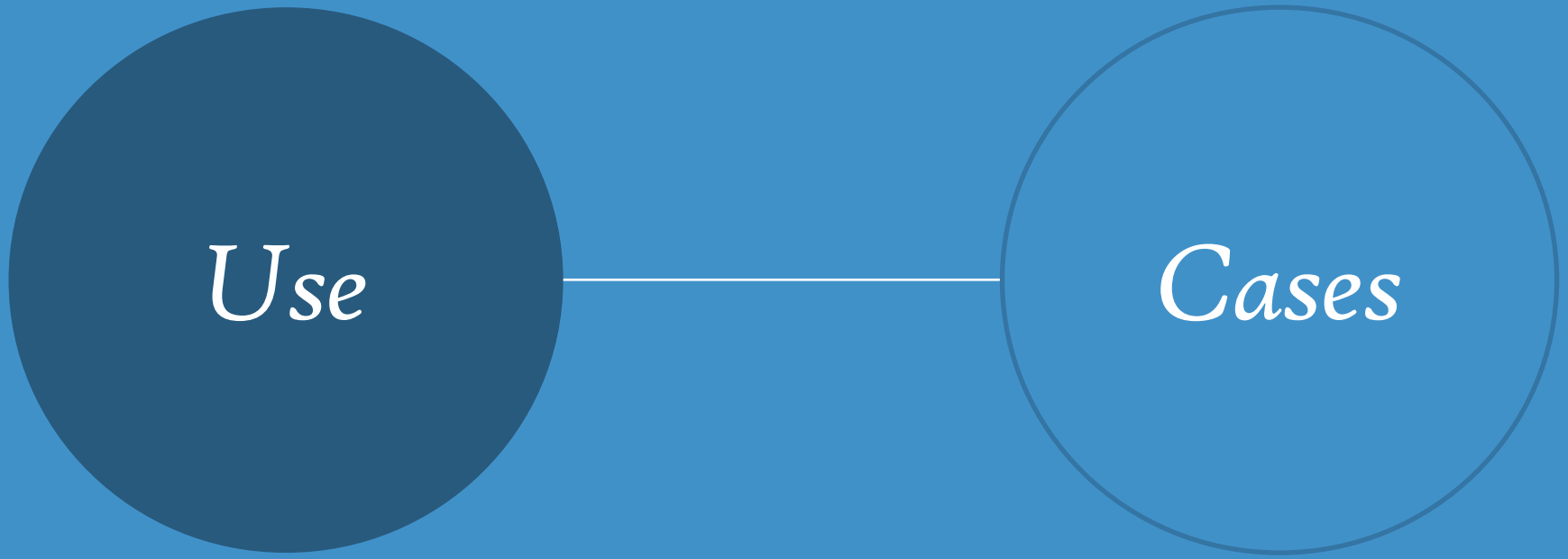
Das Thema Datenbewirtschaftung beschäftigt das Verlagswesen bereits seit Jahrhunderten. Die derzeitige Entwicklung ist nicht disruptiv, sie ist eine mehr oder weniger natürliche Weiterentwicklung. Ich glaube aber, es ist wichtig, klar zu differenzieren, was Verlagswesen bedeutet; dieser Begriff hat eine sehr große Breite. Wir als Compass-Verlag beispielsweise sind ein Unternehmen, das in seinen Verlagspublikationen inhaltlich immer schon datengetrieben war. Schon das erste Buch hatte Daten und Informationen als Inhalte. Das waren keine journalistischen oder belletristischen Inhalte, keine textuellen Inhalte im klassischen Sinn.

**Anfang des Jahrzehnts ist das Thema Big Data erstmals stark in den Blickpunkt gerückt, Daten waren plötzlich „sexy“. Wie haben Sie diese Entwicklung wahrgenommen und warum ist Ihrer Meinung nach das Datenthema plötzlich so in den Vordergrund gerückt?**

Wenn man genau hinschaut, sind 80 Prozent davon, was als Big Data verkauft wird, mehr oder weniger komplexe Datenbankabfragen. Der Sinn von Big-Data-Anwendungen sollte eigentlich sein, versteckte Muster zu erkennen; der Erkenntnisgrad von Anwendungen, die es heute am Markt gibt, ist aus meiner Sicht im Augenblick noch endenwollend. Im Bereich der verlagsdatengesteuerten Wirtschaft, in dem wir uns bewegen, sind wir bis jetzt ja mit Datenbanken sehr gut zurechtgekommen. Da stellt sich die Frage, was man mit einer Big-Data-Anwendung aus der eigenen Datenbank anderes herausholen sollte als bisher – man kennt ja die Inhalte der eigenen Datenbank, es gibt da nichts zu „entdecken“. Spannend wird es erst dann, wenn versucht wird, bestimmte Daten mit anderen Daten zu verknüpfen und zu verschneiden, um zu schauen, ob zwischen diesen Daten noch irgendwelche Korrelationen im weitesten Sinne bestehen und ob etwas daraus abgeleitet werden kann. Wobei die Frage, was abgeleitet werden kann, ja doch oft ein offenes Thema ist.

**Wo liegen also die Herausforderungen von Big Data?**

Big Data im eigentlichen Sinn dessen, was mit diesem Begriff gemeint ist, kommt erst. Bislang haben wir immer auf strukturierte Daten zugegriffen. Im Internet of Things heißt das Stichwort Mobiltelefon. Allein durch die verbauten Sensoren kommt es zu so großen Datenvolumina, dass man diese im Einzelnen gar nicht mehr anvisieren kann und man ganz andere Mechaniken zur Analyse braucht, andererseits sind diese Daten auf eine gewisse Art und Weise ungeordnet. Wenn nun Milliarden Datensätze oder Datenpunkte generiert werden und plötzlich in großer Geschwindigkeit hereinkommen, dann reichen simple Datenbankabfragen eben nicht mehr aus. Da gibt es derzeit einen massiven Engpass: Wir haben zwar die Tools, wir haben die Daten – wobei jeden Tag mehr Daten dazukommen –, aber wir haben nicht die Leute, die die richtigen Fragen stellen können.


Use
Cases

## Nature Publishing: Linked Data als Publishing-Tool

*Linked-Data-Technologien in der Content-Produktion / Mehr Kundennutzen durch LOD*

Die Nature Publishing Group (NPG), Teil der Verlagsgruppe Palgrave Macmillan, veröffentlicht rund 120 Titel, darunter renommierte wissenschaftliche Journale und Magazine wie „Scientific American" oder „Nature". Seit 2012 befasst sich die NPG mit Linked Data, experimentierte aber auch schon zuvor mit semantischen Technologien, etwa mit RSS-Feeds, verschiedenen Metadatensets, **OpenSearch** und **OpenURL**.

Im Jahr 2012 schließlich launchte die NPG eine Linked-Data-Plattform, die bibliografische Metadaten zu allen Artikeln des Verlags und deren Quellen sowie zu allen Titeln von Palgrave Macmillan enthält. Veröffentlicht wurden zu diesem Zeitpunkt rund 270 **RDF**-Tripel, die Suchfunktion wurde im April 2014 allerdings wieder stillgelegt. Dahinter stehen veränderte Zielsetzungen: Es geht zunehmend um die interne Verwendung von Linked-Data-Technologien, die nun vor allem für die Content-Produktion bzw. -Verarbeitung eingesetzt werden. Zentrale technische Schnittstelle dafür ist ein eigens realisierter „Content Hub", in dem alle Daten aggregiert werden. Während Text-Content in Form von XML-Dokumenten im Hub gespeichert ist, sind die Metadaten als Linked Data im RDF-Format zugänglich. Damit vereint die NPG im Content Hub eine praktikable Speicherlösung mit leistungsstarken Suchfunktionen.

In Zukunft sollen die an das neue Modell angepassten RDF-Datensätze wieder veröffentlicht und durchsuchbar gemacht werden. Mit der Initiative will der Verlag nach eigenen Angaben einerseits nach wie vor zu einer größer werdenden Linked-Data-Community beitragen, andererseits liegt der Fokus künftig klar auf den Bereichen Kundennutzen und interne Verwendung. Für Kunden sollen die Inhalte einfach auffindbar und interaktiv nutzbar sein; für den Verlag selbst steht die operative Nutzung des Datenmodells im Blickpunkt, um die Veröffentlichungsprozesse zu integrieren.

*Linktipps*

***nature.com Ontologies,***
*http://www.nature.com/ontologies*

## Interview: Nature Publishing Group

*„Die Welt verändert sich, und wir müssen darauf reagieren“*

*Interview mit Iain Hrynaszkiewicz, Head of Data, und Tom Scott, Director of Product Management, der Nature Publishing Group*

**Linked-Open-Data-Initiativen sind getrieben von der Idee des freien Wissens, auf der anderen Seite müssen Unternehmen auch wirtschaftlich denken. Inwiefern kann Linked Open Data Teil des Geschäftsmodells sein?**

Als Verlag ist es unsere Aufgabe, die Inhalte unserer Autorinnen und Autoren zugänglich und nutzbar zu machen. Wir wollen, dass Menschen unsere Veröffentlichungen lesen, verwenden und zitieren. Letztlich ist das die wichtigste unternehmerische Überlegung. Linked Data bietet in diesem Zusammenhang einige Vorteile. Denn die Verlagsdomäne ist am einfachsten als Graph zu beschreiben, vergleichbar mit einem Beziehungssystem. Man bedenke etwa das Netzwerk an Zitierungen, die Beziehungen von Autorinnen und Autoren zu ihren Veröffentlichungen, die Themen der Veröffentlichungen oder die **Entitäten** , die darin vorkommen. Hinzu kommt, dass wir in der Lage sein müssen, auf Veränderungen zu reagieren. Technologien wie **RDF** helfen uns dabei, agil zu bleiben, da es relativ einfach ist, **Ontologien** anzupassen oder zu erweitern. Man kann nicht länger davon ausgehen, dass man ein Modell erstellen und dann dabei bleiben kann.

**Hat sich der Einsatz von Linked Open Data für Nature Publishing bislang bewährt?**

Intern hat sich der Einsatz als erfolgreich gezeigt, da wir so eine Metadaten-Infrastruktur aufbauen konnten, die skalierbar und effizient ist. Auf dieser Basis konnten wir eine neue Plattform aufbauen, die der Website nature.com zugrunde liegt. In Hinblick auf öffentliche RDF-Daten haben wir vor kurzem unser Linked-Data-Portal aktualisiert und holen nun Feedback von verschiedenen Stakeholdern ein.

**Wie sieht die Zukunft aus? Wie geht Nature Publishing künftig mit Linked Open Data um?**

Wir planen, mehr Entitäten in unseren Wissensgraph aufzunehmen, indem wir mehr strukturierte Daten zu den Inhalten, die wir veröffentlichen, extrahieren. Zusätzlich wollen wir weitere Datensets integrieren, die von anderen Organisationen produziert und gewartet werden, wie etwa Mesh oder Wikidata.

## The Guardian: Daten, so weit das Auge reicht

*Nachrichteninhalte als Linked Data / Datenzugang zwischen Public Service und Kommerzialisierung*

Als eines der ersten großen Medienunternehmen – neben der BBC – ist die britische Tageszeitung The Guardian auf den Datenzug aufgesprungen. Während die Guardian-Redaktion im „Datablog" selbst datenjournalistisch unterwegs ist, bietet das Medienhaus mit der „Open Platform" eine Schnittstelle, über die User direkt auf Daten des Guardian zugreifen können. Seit 2010 wird dort auch Linked Data eingesetzt.

Über die „Open Platform" bietet der Guardian Zugriff auf über 1,7 Millionen Nachrichten-Items, die seit 1999 veröffentlicht wurden. Als zentrale Schnittstelle für den Zugriff fungiert die vom Guardian entwickelte „Content API", die unterschiedliche Suchmodi zulässt: nach dem gesamten Inhalt, nach Schlagworten, Rubriken, Ausgaben und Einzelitems. Die „Open Platform" bezeichnet der Guardian als „Service für die Öffentlichkeit". Zugang erhält nach dem Open-Data-Prinzip grundsätzlich jedermann, und zwar mit einem Zugangsschlüssel für Entwickler, der den Zugang zu Texten ermöglicht. Daneben vertreibt der Guardian den Zugang zu seinen Daten aber auch kommerziell: Bezahl-User haben neben Texten auch Zugriff auf Fotos, Videos und Audioinhalte.

Seit 2010 bindet die Guardian-Plattform auch externe Quellen mit ein, den Anfang machten dabei ISBN-Nummern und IDs der Musikplattform MusicBrainz. Zum einen sind solche externen Identifikatoren nun in die Berichte des Guardian – etwa Buch- oder Musikrezensionen – eingebunden, zum anderen kann über die „Content API" auch mittels dieser Suchbegriffe gesucht werden.

Anders als die meisten anderen Datenprovider stellt der Guardian seine Daten allerdings nicht in der für LOD üblichen Form von RDF-Tripeln zur Verfügung, sondern nutzt eine eigens aufgebaute Datenbankenstruktur, in die auch externe Quellen eingebunden werden können. Für Abfragen kommt eine Open-Source-Suchplattform – **Solr** – zur Anwendung.

*Linktipps*

***Open Platform,***
*http://open-platform.theguardian.com*

***Data Blog,***
*http://www.theguardian.com/data*

## Wolters Kluwer: Ein Metadaten-Ökosystem

*Vom Content-Management-System zu Semantic-Web-Strukturen / Netzwerkeffekte durch LOD / „Pure" Inhalte als Produkt*

Der Informationsdienstleister Wolters Kluwer Deutschland, der Fachpublikationen und Softwarelösungen für die Bereiche Recht, Wirtschaft und Steuern anbietet, hat im Rahmen des EU-geförderten Leuchtturmprojektes LOD2 semantische Web-Technologien in seine Redaktionsprozesse integriert. Ziel war es, Texte bzw. Teile davon leichter für verschiedene Medien zu nutzen und die Texte mit Informationen aus anderen Wissensquellen sowie Metadaten anzureichern. Dabei wurde darauf geachtet, die Datenstruktur standardisiert, wiederverwendbar und konsistent zu gestalten.

Auf technischer Ebene wurde bei Wolters Kluwer deshalb das bis dato in sich geschlossene Content-Management-System aufgebrochen. War zuvor jeder Text samt Metadaten in einer **XML**-Datei gespeichert, wurden die Textinformation und die Metadaten nun getrennt. Zur Verwaltung gibt es drei verschiedene Teilsysteme: ein Content-Management-System, in dem Texte als XML-Dateien gespeichert sind; ein kontrolliertes Vokabular für Metadaten im **SKOS**-Format und weitere Metadaten sowie Informationen über die Struktur von Dokumenten im **RDF**-Format.

Diese Trennung aus Text- und Metainformation ermöglichte es, „sukzessive die alte Redaktionsinfrastruktur zu modernisieren und neue Anforderungen aus den Produkten wie Personalisierung oder semantische Suche zu adressieren", schreiben die Content-Strategen von Wolters Kluwer. Durch den Aufbau eines ganzheitlichen Ökosystems für Metadaten mit Schnittstellen zu anderen Quellen seien wesentliche Netzwerkeffekte erzielbar. Das Unternehmen sieht Linked Data als Teil des Geschäftsmodells: Durch die koordinierte Zusammenarbeit zwischen Wissens-Lieferanten und - Konsumentinnen und Konsumenten kann der rein wissenschaftliche Nutzen auch ökonomisch verwertet werden, schließt man bei Wolters Kluwer.

Bis dato hat das Unternehmen zwei **Thesauri**, den WKD Arbeitsrechtthesaurus und den WKD Gerichtsthesaurus, im RDF-Format sowie entsprechende Abfragepunkte veröffentlicht.

*Linktipps*

***Projekt LOD2,***
*http://lod2.eu/*

***WKD Arbeitsrechtthesaurus,***
*http://vocabulary.wolterskluwer.de/PoolParty/sparql/arbeitsrecht*

***WKD Gerichtsthesaurus,***
*http://vocabulary.wolterskluwer.de/PoolParty/sparql/court*

## Interview Wolters Kluwer

*„Das Engagement hat sich bewährt“*

*Interview mit Christian Dirschl, Content Architect und Leiter Content Strategy bei Wolters Kluwer Deutschland*

**Linked-Open-Data-Bestrebungen sind getrieben von der Idee des freien Wissens, auf der anderen Seite müssen Unternehmen auch wirtschaftlich denken. Inwiefern kann Linked Open Data Teil des Geschäftsmodells sein?**

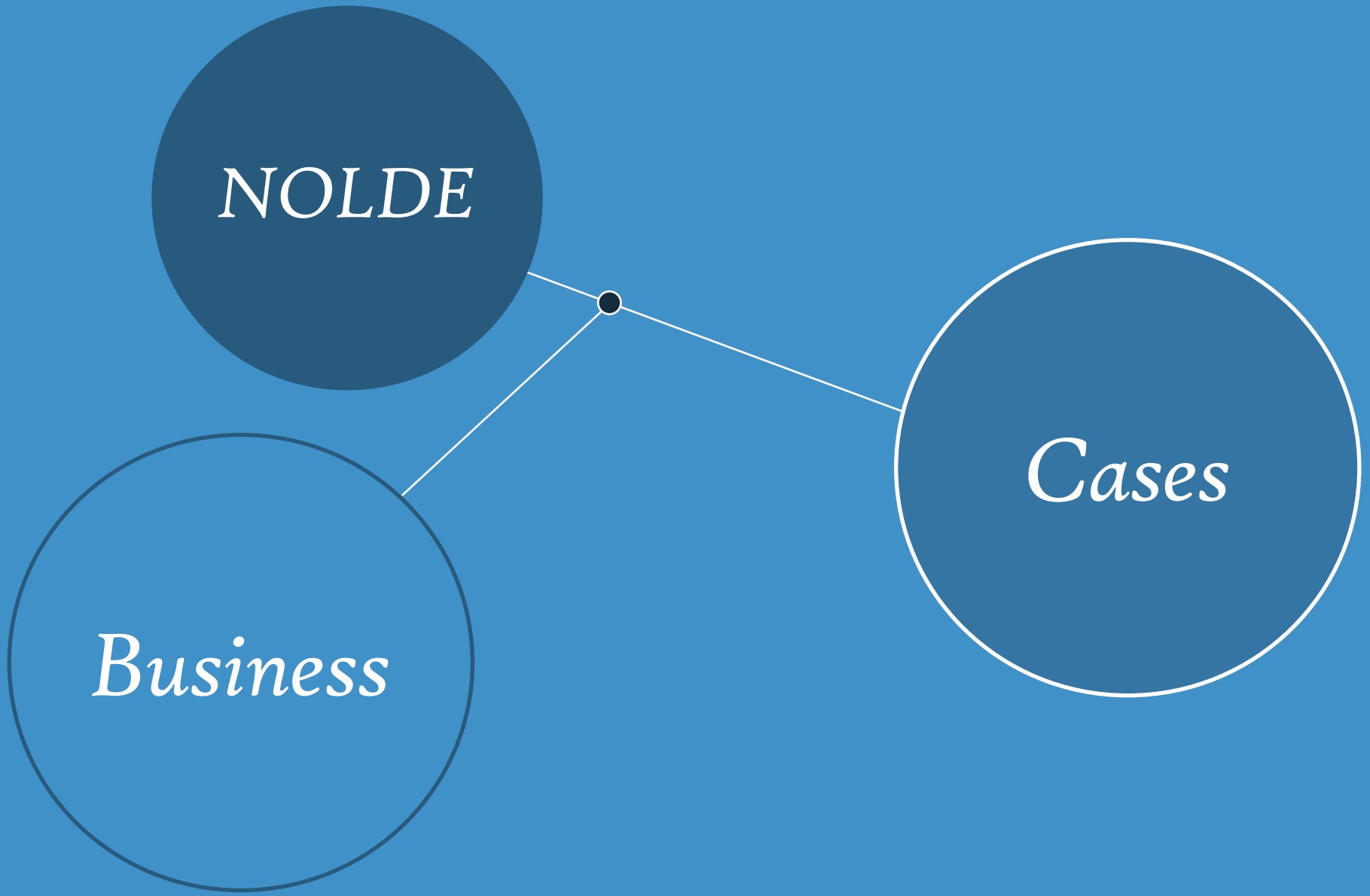
Auf der einen Seite ist man Data Consumer, kann also seine eigenen Daten anreichern, ohne selbst den Aufwand betreiben zu müssen. Auf der anderen Seite können wir uns mit Linked Open Data auch als Unternehmen weiterentwickeln. Wir kommen ursprünglich aus der Publishing-Umgebung, hatten immer fertige Produkte, zwischen zwei Buchdeckeln oder auf einer CD, für unsere Kunden. Mittlerweile stellen wir Content auch unabhängig von konkreten Produkten zur Verfügung, durch interne Datenbanken oder APIs. So können unsere Kunden den Content bei sich integrieren, wie sie es wollen, und nicht, wie wir es wollen.

**Hat sich der Einsatz von Linked Open Data für Wolters Kluwer Deutschland bislang bewährt?**

Für Linked Open Data ist auch der Aufbau einer technologischen Infrastruktur nötig, die in der Folge auch für interne Daten und Prozesse nutzbar ist. Wir haben Teile unserer Technologie auf Semantic-Web-Standards umgestellt. Diese Technologien nutzen wir bereits operativ und können damit Dinge machen, die man mit einem klassischen Content-Management-System nicht tun kann. Insofern hat sich unser Engagement als Gesamtkonstrukt sicherlich auch finanziell schon bewährt. Was aber die eigentliche Datennutzung betrifft, so sind die vorhandenen Daten, die für uns im juristischen Bereich relevant wären, nicht wirklich nutzbar.

**Wie sieht die Zukunft aus? Wie geht Wolters Kluwer Deutschland künftig mit Linked Open Data um?**

Wir nehmen aktiv an verschiedensten Bestrebungen teil, um die derzeit existierenden Probleme, etwa bezüglich Datenqualität oder Lizenzierung, mitzubeheben. Parallel dazu versuchen wir, interne Prozesse weiter zu optimieren, um Linked-Open-Data-Informationsflüsse, sobald sie wirklich kommen, besser verarbeiten zu können.


NOLDE
Cases
Business

## Monopol Verlag: Austrian Music Monitor

*Ressort- und plattformübergreifende Redaktionsarbeit unterstützen / Automatischer Aufbau von Wissensbasen / Teil der LOD Cloud werden*

Der Monopol Verlag nutzt Linked Open Data gemeinsam mit seinem bestehenden Datenbestand aus dem Musikbereich. Der Verlag verfügt über eine umfangreiche Datenbank zu Musikrezensionen über österreichische Musikschaffende. Diese wurden im Rahmen eines Pilotprojektes mit Daten aus der LOD-Cloud sowie Analysedaten aus dem Social-Media-Monitoring angereichert. Ziel des Verlages ist der Aufbau eines Portals mit einer offenen Datenschnittstelle, welche die aggregierten Daten der Öffentlichkeit zugänglich macht. Diese Prinzip soll in Folge auch auf andere Themenbereiche ausgeweitet werden.

In der ersten NOLDE-Projektphase lag der Fokus darauf, die Künstlerinformationen zu sammeln und relevante Daten aus der LOD-Cloud (z.B. BBC Music, MusicBrainz) zu integrieren; neben der Ergänzung der bestehenden Daten wurden diese auch zur Anreicherung eigener redaktioneller Inhalte verwendet. Zusätzlich wurden Echtzeit-Daten aus sozialen Medien wie Facebook, Twitter, SoundCloud, LastFM und YouTube hauseigenen Social-Media-Monitoring-System, Social Media Ranking gesammelt. Aktuell wird an einem Portal gearbeitet, dass diese Daten in enzyklopädischer Form der Öffentlichkeit zugänglich macht. Zusätzlich wird ein kuratierter Auszug der aggregierten Daten mittels einer offenen Programmierschnittstelle verfügbar sein.

Als erstes Teilprodukt entsteht eine **Rich Content**-Enzyklopädie des zeitgenössischen österreichischen Musikschaffens (vergleichbar mit http://linkedjazz.org/). Dieses Tool soll in Folge soziale Daten von Band-Pages und deren Fans und für Live-Services und Trendanalysen sammeln. Weiters soll daraus ein B2B-Informations- und Vermarktungsservice für die österreichische Musikwirtschaft entstehen, das Management- und Vermarktungsprozesse für Labels, Agenten, Rechteinhaber und Band-Management optimiert. Dies erfolgt durch eine systematische Analyse der Daten, die wichtige Kennzahlen – wie Fans, Followers, Interaktionen oder Geo-Locations – für eine bessere Entscheidungsgrundlage in der Musikwirtschaft generiert. Die dabei erhobenen Daten sollen u.a. in der Werbeoptimierung Verwendung finden, oder auch für die ressortübergreifende Kombination von Lifestyle-Themen im Magazin-Portfolio des Monopol Verlags.

# Interview: Monopol-Verlag

## *„Wir haben ein doppeltes Interesse an Linked Data“*

*Interview mit Martin Mühl, Geschäftsführer beim Monopol Verlag.*

**Was erwartet sich der Monopol-Verlag vom Einsatz von Linked Data?**

Wir erhoffen uns prinzipiell eine positive Entwicklung auf der eigenen Website, einerseits in Sachen Zugriffe, andererseits durch Verlinkungen. Im Idealfall lassen sich auch von uns verschiedene Services, die wir anbieten, besser miteinander verknüpfen, sodass die User besser von einem Service zum anderen gelenkt werden.

**Wo kommt Linked Data beim Monopol-Verlag zum Einsatz?**

Der zentrale Teil wird unsere eigene redaktionelle Website sein. Wir haben aber unterschiedliche zusätzliche Services, etwa im Social-Media-Bereich. Ein Projekt, das wir ursprünglich gemeinsam mit einer Schwesterfirma entwickelt haben, ist „Music Meta“. Dabei handelt es sich um einen Social-Media-Radar, der für verschiedene Unternehmen Aktivitäten, Zugriffe, Erfolge, Nicht-Erfolge auf Facebook, Foursquare, Twitter, Google+ etc. misst. Wir haben hier ein eigenes Service im Musikbereich, das österreichische Bands, Venues etc. monitort und aussagt, welche Trends sich ergeben. Das wollen wir weiter verbinden und mit unserer redaktionellen Arbeit verknüpfen.

**Wie passt Linked Data in die Entwicklungsstrategie des Monopol-Verlags?**

Wir haben ein doppeltes Interesse an Linked Data: Zum einen berichten wir mit unserer Medienmarke „The Gap“ immer wieder über technische Entwicklungen, die unser Umfeld und unsere Leserinnen und Leser betreffen. Zum anderen hoffen wir auch, dass unsere eigenen Angebote und Services von Linked Data profitieren.

**Wie sieht die Zukunft aus? Wie geht der Monopol-Verlag künftig mit Linked Open Data um?**

Wir haben uns innerhalb des NOLDE-Projekts darauf geeinigt, mit Linked Data im Musikbereich zu starten und auf unserer neuen Website hauptsächlich im Musikbereich zu vernetzten. Wenn das zu positiven Ergebnissen führt, dann wollen wir das Projekt auf andere redaktionelle Bereiche ausweiten. Außerdem gehen wir davon aus, dass sich zusätzliche Angebote im B2C-, wie auch im B2B-Bereich daraus ergeben.

## *Verlag des Österreichischen Gewerkschaftsbundes: Arbeitsrecht verständlich machen*

*Content-Aggregation über mehrere CMS unterstützen / Adaptive Filtermechanismen entwickeln / Syndizierung mit Content-Partnern erleichtern*

Hilfe zur Selbsthilfe: Der Verlag des Österreichischen Gewerkschaftsbundes (ÖGB) ist spezialisiert auf des Thema Arbeitsrecht und verfügt über einen großen Bestand an Kollektivverträgen, relevanten Gesetzestexten, Verordnungen, EU-Richtlinien, Kommentaren, Monografien, Aufsätzen und Ratgebern. Im Rahmen eines Pilotprojekts wird versucht, diese Datenmengen mithilfe von Linked-Data-Technologien aufzubereiten und als Wissensbasis für Interessierte – quasi als Ermächtigung zur Selbsthilfe – auszubauen.

In einem ersten Schritt soll das System Empfehlungen für die Suche nach dem relevanten Kollektivvertrag (aus derzeit insgesamt 600 liefern), außerdem soll es Kurzdarstellungen der wichtigsten und meistgefragten Inhalte geben. Das KV-System soll aber auch in ein übergreifendes arbeitsrechtliches Informationssystem eingebettet werden, das neben Kollektivverträgen auch andere Rechtsquellen heranzieht. So sollen darüber hinaus die technischen Voraussetzungen für eine automatische Verrechnung und Cross-Lizensierung geschaffen werden.

Über leicht bedienbare Schnittstellen (APIs) sollen die Daten auch Partnern und der Öffentlichkeit zugänglich gemacht werden. So soll es beispielsweise möglich sein, Zeitreihenanalysen zu machen oder detaillierte Lohn- und Gehaltstafeln anzusehen. Gleichzeitig sollen auch weitere externe Quellen, wie zum Beispiel das Rechtsinformationssystem des Bundes (RIS), eingebunden und relevante Daten aus dem gesamten deutschsprachigen Raum ergänzt werden.

Nach Projektende von NOLDE plant der ÖGB-Verlag, das System neben dem Arbeitsrecht auf weitere Domänen wie Arbeitnehmerschutz und Sozialrecht auszudehnen.

# Interview: ÖGB-Verlag

*„Open Data passt sehr gut zu unserer Philosophie“*

*Interview mit Christian Wachter, zuständig für Wissensmanagement im ÖGB-Verlag.*

**Was erwartet sich der ÖGB-Verlag vom Einsatz von Linked Data?**

Wir wollen unsere Inhalte anderen besser zur Verfügung stellen und an den Inhalten anderer besser partizipieren können. Der Austausch von Inhalten und von Wissen soll also effektiver werden.

**Wo kommt Linked Data beim ÖGB-Verlag zum Einsatz?**

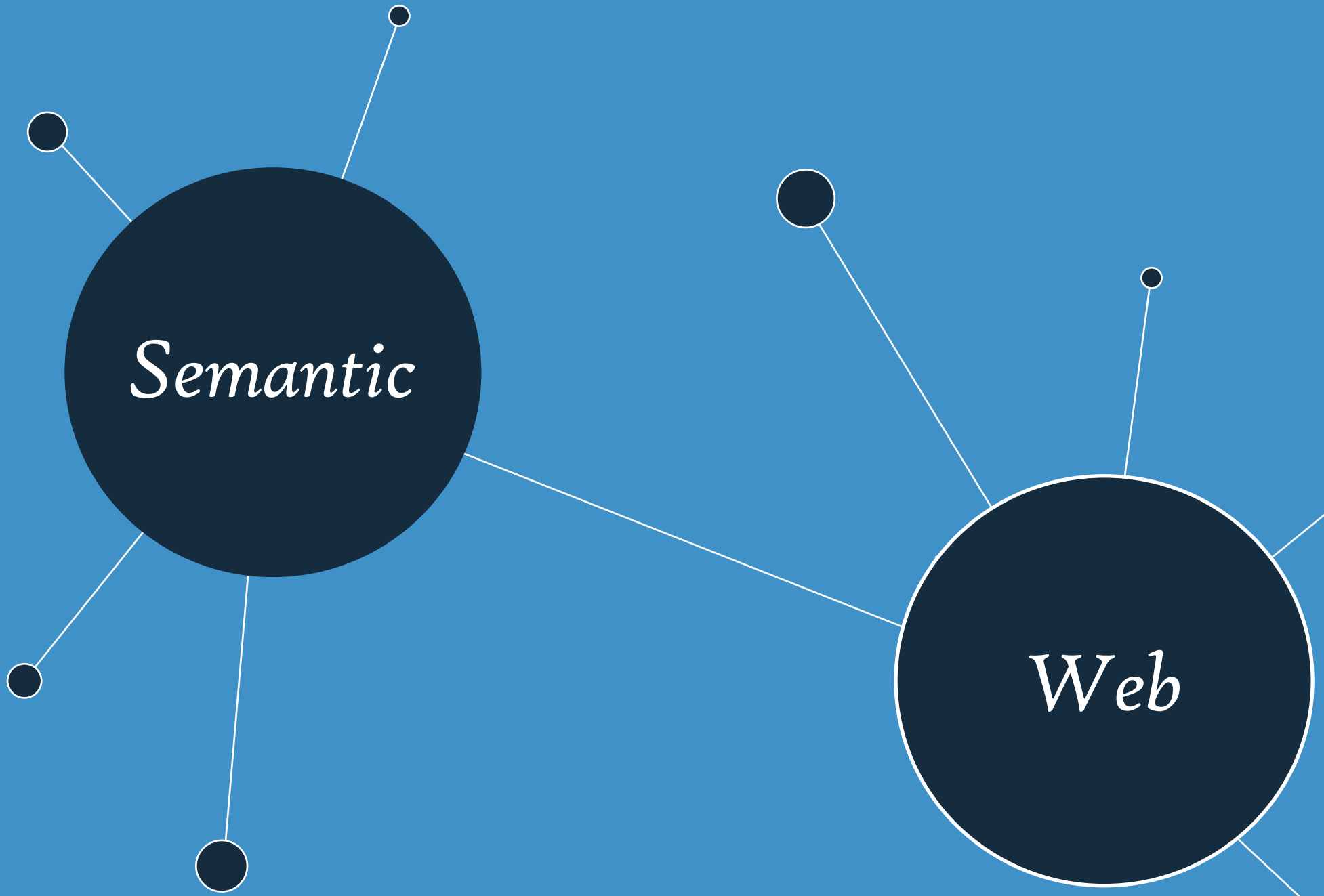
Wir sind dabei, einen Thesaurus für Arbeitsrecht zu entwickeln. Hier könnten Linked-Data-Anwendungen zum Einsatz kommen. Auch, wenn das der Benutzer nicht direkt sehen kann: So kann die Suche in verschiedenen Webapplikationen durch Linked Data besser funktionieren. Des Weiteren können wir den Thesaurus in Form von Open Data anderen zur Verfügung stellen und gemeinsam weiterentwickeln. Umgekehrt können wir genauso Datenquellen von außen anzapfen, etwa das Rechtsinformationssystem des Bundes, RIS, und über dieses Wissensmodell mit unseren Inhalten verknüpfen.

**Wie passt Linked Data in die Entwicklungsstrategie des ÖGB-Verlags?**

Im Unterschied zu anderen Verlagen ist der ÖGB-Verlag ein politisches Projekt, da sein Eigentümer der Österreichische Gewerkschaftsbund ist; auch die Arbeiterkammern stehen dem sehr nahe. Für diese Gruppen arbeiten wir. Bei Linked Open Data geht es darum, dass man Inhalte nicht für sich behält, sondern diese der Community offen zur Verfügung stellt. Es geht also nicht nur um technische Aspekte, sondern auch um den gesellschaftlichen Aspekt, dass Wissen mehr Nutzen bringt, wenn man es gemeinsam nutzt und weiterentwickelt. Insofern passt der Open-Data-Ansatz – die Vernetzung mit anderen – sehr gut zu unserer Philosophie.

**Wie sieht die Zukunft aus? Was bringt dem ÖGB-Verlag Linked Data künftig?**

Wir wollen durch den Einsatz von Linked Data zukünftig im Bereich Recht, speziell im Bereich des Arbeitsrechts, besser mit anderen Institutionen zusammenarbeiten. Das sind unsere Partner in der Verlagslandschaft genauso wie wissenschaftliche Institutionen oder die Sozialpartner.


Semantic
Web

## Das Semantische Web I: Die Grundprinzipien

*Von der Utopie des semantischen Webs zur Realität / Vier Prinzipien, die Linked Data prägen*

Die Idee von Linked bzw. Linked Open Data ist in ihren Grundsätzen nicht neu – genauer gesagt ist sie genauso alt wie das World Wide Web selbst. Sie fußt auf Sir Tim Berners Lees Vorstellung des semantischen Webs, das der Erfinder des World Wide Web schon in seinen ursprünglichen Skizzen des WWW mitgedacht hatte.

Der Ansatz ist einleuchtend: Im Web stehen unzählige Multimedia-Dokumente, Anwendungen und große Mengen an strukturierten Daten bereit, die ein großes Potenzial zur (Weiter-)Nutzung haben. Doch der Zugriff auf Dokumente und darin enthaltene Daten gestaltet sich schwierig: einerseits aufgrund verschiedener Kodierungsformate (etwa HTML, PDF oder proprietäre Formate), andererseits durch die Verwendung von Datenbanken und Content-Management-Systemen, die den Zugriff auf die dort gespeicherten Daten oft unmöglich machen. Auch Schnittstellen (APIs), mit denen einzelne Anbieter Zugriff auf ihre Daten gewähren, stellen keine wirkliche Lösung dar. Die Nutzung unterliegt bekanntermaßen technischen, wie auch vom Anbieter vorgegebenen Einschränkungen, außerdem sind APIs nicht anbieterübergreifend nutzbar.

Es gibt also ungeheure Mengen an potenziell nutzbaren Daten, die aber in geschlossenen Datensilos feststecken oder, sofern sie verfügbar sind, nicht standardisiert nutzbar sind. Linked Data kann als Lösungsansatz für dieses Problem verstanden werden. Er macht sich bestehende Technologien zunutze, die das World Wide Web Consortium (W3C) als Semantic-Web-Technologien standardisiert hat. Diese spiegeln sich auch in den vier Richtlinien Tim Berners Lees wider, die heute als Linked-Data-Prinzipien bekannt sind:

1. Verwende URIs (Uniform Resource Identifier), um Dinge, Gegenstände und Konzepte zu identifizieren und zu addressieren.
2. Verwende URIs, die via HTTP (Hypertext Transfer Protocol) aufgelöst (dereferenziert) werden, damit sie im Web nachgeschlagen werden können.
3. Wird eine URI im Web aufgegriffen, dann halte nützliche Informationen dazu bereit und verwende die W3C-Standards RDF(S) und SPARQL.
4. Verknüpfe deine Daten mit anderen URIs, damit Benutzer noch mehr im Web of Data entdecken können.

Was auf den ersten Blick kompliziert klingt, klärt sich bei genauerer Betrachtung der Technik, die hinter dem Web of Data steht. Denn die Technologien, die für Linked (Open) Data verwendet werden, sind vielfach dieselben, die wir aus dem World Wide Web bereits kennen.

# Das Semantische Web II: Wie das „Linked“ in Linked Data kommt

*Wie Daten im Web of Data beschrieben werden / Wie Datenbanken zwischen Mensch und Maschine unterscheiden / Wie aus Data „Linked Data“ wird*

Am Anfang war der Link: Die Idee von Hyperlinks gibt es so lang wie das Internet selbst. Genauer gesagt ist es dieses Referenzierungsschema, das das World Wide Web überhaupt erst zum Netz macht. Der gleichen Technik bedient sich auch **Linked Open Data**. Aber von Anfang an.

Jede Website, jedes Dokument im World Wide Web hat eine eindeutig zuordenbare Adresse. So erreicht man etwa die österreichische Version der Suchmaschine Google unter der Adresse www.google.at. Diese Adressen werden als **URL**s (**Uniform Resource Locators**) bezeichnet und geben an, wo im Netz sich die gesuchte Seite befindet. Ändert sich der Speicherort der Seite, ändert sich auch die URL, mit der sie aufgerufen werden kann. Daneben gibt es sogennante **URN**s (**Uniform Resource Names**). Das sind dauerhafte Namen, die vom Ort bzw. der Adresse eines Dokuments im Web unabhängig sind und gesondert verwaltet werden können.

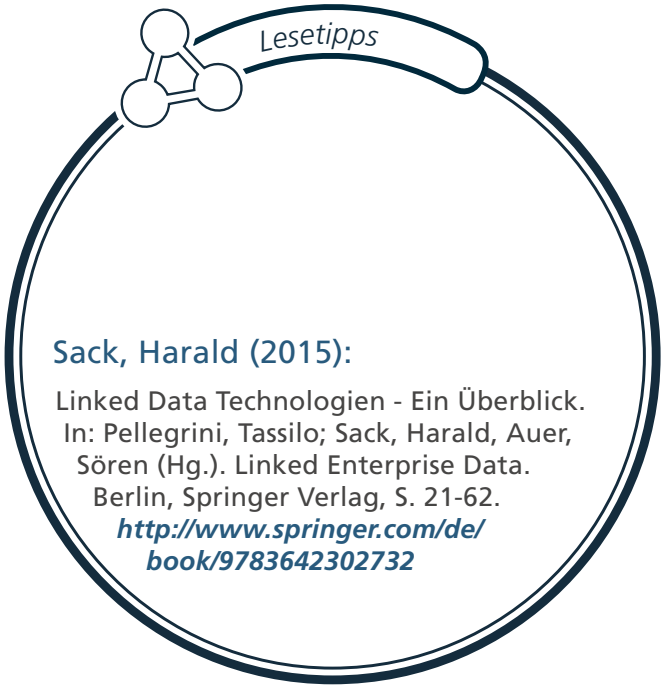
Für Linked Open Data macht man sich diese beiden Schemata zunutze und verwendet **URI**s (**Uniform Resource Identifiers**), eine Art Kombination aus URLs und URNs. So ist es möglich, nicht nur Dokumente wie Websites miteinander zu verknüpfen, sondern alle möglichen Formen von Daten. Neben Objekten im Web können so auch Informationen über Objekte in der realen Welt (sogenannte Metadaten) miteinander verknüpft werden.

Als „Sprache“ für die Beschreibung von Objekten dient bei Linked-Data-Ressourcen der Formatstandard **RDF** (**Resource Description Format**). Alle Aussagen über ein Objekt werden dabei in einem Tripel-Schema gemacht, das einem einfachen Satz in einer menschlichen Sprache ähnelt. Jedes Tripel besteht aus „Subject“, „Property“ und „Object“. Damit lassen sich etwa Aussagen tätigen wie: Dokument.html (Subject) hat den Autor (Property) Thomas (Object). Jeder der drei Bestandteile wird in Form eines URIs dargestellt.

RDF-Tripel strukturieren die vorhandenen Informationen. Um dem gesamten Datenmodell Bedeutung zu geben, müssen aber auch Aussagen über die Beziehungen der Tripel untereinander gemacht werden können. Dazu gibt es die Modellierungssprache **RDFS** (**RDF Schema Description Language**). Sie ermöglicht es, gleiche bzw. ähnliche Elemente in Klassen zu gruppieren, Instanzen dieser Klassen zu definieren oder Beziehungen zwischen den einzelnen Klassen festzulegen. Wenn darüber hinaus noch komplexere Strukturen abgebildet werden sollen, kommen sogenannte Ontologien zum Einsatz. Eine vom Webkonsortium W3C festgelegte **Ontologie** ist **OWL** (**Web Ontology Language**).

Auch die Übertragung von Linked Data funktioniert nicht anders, als wir es aus dem Web gewohnt sind: über das Zugriffsprotokoll HTTP (Hypertext Transfer Protocol). Das funktioniert, weil man sich beim Abrufen von Webinhalten ganz einfach als Mensch oder Maschine, sprich ein abfragendes Programm, identifizieren kann – wenn wir beispielsweise im Web surfen, übernimmt unser Browser diese Aufgabe für uns. Wenn nun ein Programm Zugriff auf RDF-Daten benötigt, sendet es diese Information bei der HTTP-Anfrage mit, und die Daten werden in maschinenlesbarem Format ausgeliefert und können so direkt weiterverarbeitet werden.

Für die Abfrage von RDF-basierten Inhalten gibt es ebenfalls einen vom Webkonsortium W3C herausgegebenen Standard. In Anlehnung an die Datenbanken-Abfragesprache SQL nennt sich die Abfragesprache für RDF-Inhalte **SPARQL** (SPARQL Protocol and RDF Query Language). Vereinfacht gesagt werden vor der Abfrage Muster aus RDF-Tripeln festgelegt, sodass passende Tripel während der Suche aus dem Datenbestand herausgefiltert werden.



## *PoolParty* - Ein Tool der Semantic Web Company

*„Mit PoolParty kann man relativ simpel seine Welt abbilden“*

*Interview mit Florian Huber, Technical Consultant der Semantiv Web Company*

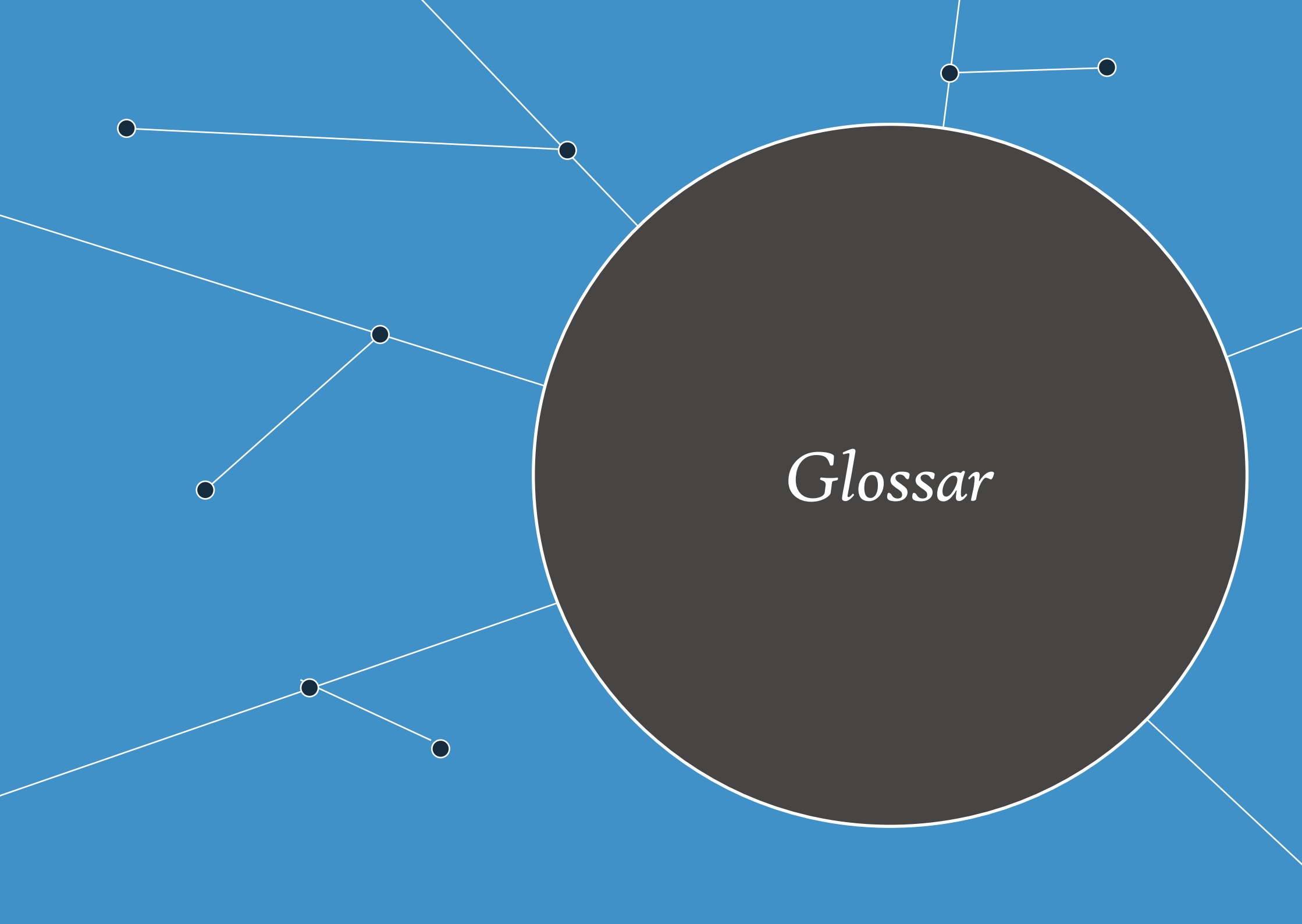
Was kann PoolParty? „Erstaunlich viel“, sagt Florian Huber. Er ist „technical consultant“ bei der Semantic Web Company, ist mit der Entwicklung des Tools beschäftigt und muss auch Kundinnen und Kunden immer wieder erklären, was dieses Programm kann. Die kurze Antwort: Man kann damit Unternehmensdaten verwalten und analysierbar machen.

„Eine Organisation hat einen Datenbestand“, so Huber, „üblicherweise in Form von Datenbanken. Ein Problem wird es, wenn es darum geht, über den Tellerrand hinaus zu schauen: Wie stehe ich da im Vergleich zu einer Partnerorganisation? Kann ich meine Daten mit öffentlichen Daten verknüpfen?“

Hier kommt PoolParty ins Spiel. Damit können Verlage ihre Daten verwalten und mit Daten aus der „Linked Data Cloud“ verknüpfen. Der Kern von PoolParty ist der Thesaurus-Manager. „Er erlaubt, dass man an einer Stelle seine Begriffe sammelt, hierarchisch strukturiert und auch Beziehungen zwischen diesen Begriffen herstellt. PoolParty stellt dafür das Grundgerüst zur Verfügung“, sagt Huber. Wie kann das genau aussehen? Erklären wir es anhand eines Beispiels, das so in der Musiker-Datenbank eines Verlages auftauchen könnte.

„Rainhard Fendrich ist eine reale Person. Sein Name ist eine Entität, ein Konzept, das mit anderen Konzepten in Verbindung steht. Sein Geburtsort ist ein weiteres Konzept. Schon haben wir eine Beziehung zwischen dem Konzept Rainhard Fendrich und seinem Geburtsort. Dann hat er verschiedene Lieder gesungen. Man kann jeden Song als einzelnes Konzept definieren.“ Da PoolParty auf dem Standard SKOS (Simple Knowledge Organisation System) aufbaut, sei es relativ simpel, damit seine Welt abzubilden.

Die nächste Stufe ist der Extraktor. „Sobald man sein Vokabular zusammengetragen hat, kann man zum Beispiel sein Archiv beschlagworten. Ein Extraktor zerlegt die Texte und schaut, dass die Zerlegung der einzelnen Wörter mit den Begriffen im Thesaurus zusammenpasst“, so Huber. „Man nimmt einen Text her und setzt die Wörter mit den zuvor definierten Konzepten in Beziehung. Die Idee ist, dass man dazu eine semantische Suche anbietet. Da wird wieder das Wissen vom Thesaurus genutzt.“ Auf diese Weise könnte man etwa gezielt nach allen steirischen Musikern suchen, die auf Englisch singen und mit Punk zu tun haben. PoolParty geht aber noch weiter: Man kann zusätzlich (maschinenlesbare) Daten aus anderen Quellen abholen und mit den eigenen Daten verknüpfen. Huber: „Da gibt es viel Material. Da könnte man Fotos übernehmen, die Biografie, Diskografie.“


Glossar

# GLOSSAR

**Big Data** bezeichnet Datenmengen, die zu groß oder zu komplex sind oder sich zu schnell ändern, um sie mit manuellen und klassischen Methoden der Datenverarbeitung auszuwerten. Ergänzend wird mit Big Data auch oft der Komplex der Technologien beschrieben, die zum Sammeln und Auswerten dieser Datenmengen verwendet werden. [1]

**Creative Commons** ist eine gemeinnützige Organisation, die 2001 in den USA gegründet wurde. Sie veröffentlicht verschiedene Standard-Lizenzverträge, mit denen ein Autor der Öffentlichkeit auf einfache Weise Nutzungsrechte an seinen Werken einräumen kann. Diese Lizenzen sind nicht auf einen einzelnen Werkstyp zugeschnitten, sondern für beliebige Werke anwendbar, die unter das Urheberrecht fallen. [2]

**Entitäten** werden in Auszeichnungssprachen (Markup Languages) verwendet, um wiederkehrende Informationseinheiten zu standardisieren. [3]

**Interoperabilität** ist die Fähigkeit zur Zusammenarbeit von verschiedenen Systemen, Techniken oder Organisationen. Dazu ist in der Regel die Einhaltung gemeinsamer Standards notwendig. [4]

**Linked Data/Linked Open Data** bezeichnet im World Wide Web frei verfügbare Daten, die per Uniform Resource Identifier (URI) identifiziert sind und darüber direkt per HTTP abgerufen werden können und ebenfalls per URI auf andere Daten verweisen. Idealerweise werden zur Kodierung und Verlinkung der Daten offene Standards verwendet. Dort, wo der Schwerpunkt weniger auf der freien Nutzbarkeit der Daten liegt, ist auch die Bezeichnung Linked Data üblich. [5]

**LOD-Cloud** bezeichnet das weltweite Netz aus Daten, die durch Linked-Data-Standards miteinander verknüpft sind. Als LOD-Cloud wird auch eine Visualisierung bezeichnet, die alle Linked-Data-Quellen in einem Cloud-Diagramm darstellt. [6]

**ODRL, Open Digital Rights Language**, ist eine XML-basierte Standardsprache zur Rechtebeschreibung (Rights Expression Language, REL), mit der Befugnisse, Verbote, Verpflichtungen und Geltendmachungen für digitale Inhalte ausgedrückt werden können. [7]

**Ontologien** sind meist sprachlich gefasste und formal geordnete Darstellungen einer Menge von Begrifflichkeiten und der zwischen ihnen bestehenden Beziehungen in einem bestimmten Gegenstandsbereich. Sie werden dazu genutzt, Wissen in digitalisierter und formaler Form zwischen Anwendungsprogrammen und Diensten auszutauschen. [8]

**Open Data** bedeutet die freie Verfügbar- und Nutzbarkeit von – meist öffentlichen – Daten. Sie beruht auf der Annahme, dass vorteilhafte Entwicklungen unterstützt werden, wenn adressatengerecht und benutzerfreundlich aufbereitete Informationen öffentlich zugänglich gemacht werden und damit mehr Transparenz und Zusammenarbeit ermöglichen. Dazu verwenden die Ersteller Lizenzmodelle, die auf das Urheberrecht, Patente oder andere proprietäre Rechte weitgehend verzichten. [9]

**Open Data Commons** ist ein Projekt der Open Knowledge Foundation (OKF), das rechtliche Lösungen für freie Daten bereitstellt. Es pflegt eine Reihe von Lizenzen für freie Datenbanken. [10]

**OpenSearch** ist eine auf XML basierende Sammlung von Techniken, die es ermöglicht, Suchergebnisse von Suchmaschinen und Websites in einem standardisierten und maschinenlesbaren Format auszugeben. [11]

**OpenURL** ist ein Standard zur Angabe von Metadaten in einer URL, um unabhängig vom aktuellen Speicherort auf elektronische Dokumente zu verlinken. [12]

**OWL, Web Ontology Language**, ist eine Spezifikation des World Wide Web Consortiums (W3C), um Ontologien anhand einer formalen Beschreibungssprache erstellen, publizieren und verteilen zu können. Es geht darum, Termini einer Domäne und deren Beziehungen formal so zu beschreiben, dass auch Software (z.B. Agenten) die Bedeutung verarbeiten kann. [13]

**RDF, Resource Description Framework**, bezeichnet eine technische Herangehensweise im Internet zur Formulierung logischer Aussagen über beliebige Dinge (Ressourcen). Im RDF-Modell besteht jede Aussage aus den drei Einheiten Subjekt, Prädikat und Objekt, wobei eine Ressource als Subjekt mit einer anderen Ressource oder lediglich einem Wert als Objekt näher beschrieben wird. Mit einer weiteren Ressource als Prädikat bilden diese drei Einheiten ein Tripel. [14]

**RDFS, Resource Description Framework Schema**, stellt ein Vokabular zur Verfügung, mit dessen Hilfe eine bestimmte Anwendungsdomäne modelliert werden kann. Außerdem können die in der Domäne vorkommenden Ressourcen, ihre Eigenschaften und Relationen untereinander durch RDFS repräsentiert werden. Man kann also mit RDFS einfache Ontologien formalisieren. [15]

**Rich Content** bezeichnet Inhalte, die optisch und akustisch durch beispielsweise Video, Audio und Animation angereichert werden. [16]

**Semantic Web** erweitert das Web, um Daten zwischen Rechnern einfacher austauschbar und für sie einfacher verarbeitbar zu machen. Diese zusätzlichen Informationen explizieren die sonst nur unstrukturiert vorkommenden Daten. Zur Realisierung dienen Standards zur Veröffentlichung und Nutzung maschinenlesbarer Daten, insbesondere RDF. [17]



**SKOS, Simple Knowlegde Organization System**, ist eine auf dem RDF bzw. RDFS basierende formale Sprache zur Kodierung von Dokumentationssprachen. Mit SKOS soll die einfache Veröffentlichung und Kombination kontrollierter, strukturierter und maschinenlesbarer Vokabulare für das semantische Web ermöglicht werden. [18]

**Solr** ist eine Open-Source-Suchplattform, die auf der Programmbibliothek Apache Lucene zur Volltextsuche basiert. Solr kommuniziert über das Hypertext Transfer Protocol (HTTP). [19]

**SPARQL, SPARQL Protocol and RDF Query Language**, ist eine graphenbasierte Abfragesprache für RDF. [20]

**Thesaurus bzw. Wortnetz** ist ein kontrolliertes Vokabular, dessen Begriffe durch Relationen miteinander verbunden sind. [21]

**URI, Uniform Resource Identifier**, ist ein Identifikator und besteht aus einer Zeichenfolge, die zur Identifizierung einer abstrakten oder physischen Ressource dient. URIs werden zur Bezeichnung von Ressourcen im Web (wie Webseiten, Dateien, Aufruf von Webservices, aber auch z.B. E-Mail-Empfängern) eingesetzt. [22]

**URL, Uniform Resource Locator**, identifiziert und lokalisiert eine Ressource, wie z.B. eine Website, über die zu verwendende Zugriffsmethode (z.B. das verwendete Netzwerkprotokoll wie HTTP oder FTP) und den Ort der Ressource in Computernetzwerken. [23]

**URN, Uniform Resource Name**, ist ein Uniform Resource Identifier (URI), der als dauerhafter, ortsunabhängiger Bezeichner für eine Ressource dient. Anders gesagt werden URNs dazu benutzt, Ressourcen eindeutige und dauerhaft gültige Namen zu geben, um sie somit eindeutig identifizieren zu können. [24]

**Wissensgraph** ist eine Wissensdatenbank, die dazu verwendet wird, Suchergebnisse mit semantischen Suchinformationen aus verschiedenen Quellen anzureichern. [25]

**Web of Data** ist eine synonyme Bezeichnung für das semantische Web. [26]

**XML, Extensible Markup Language**, ist eine Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten in Form von Textdateien. XML wird u.a. für den plattform- und implementationsunabhängigen Austausch von Daten zwischen Computersystemen eingesetzt, insbesondere über das Internet. [27]

## Quelle: Wikipedia, CC BY-SA 3.0

**1-6, 8, 11, 13-24, 27**: Deutsche Wikipedia, angepasst bzw. gekürzt
**10, 12**: Deutsche Wikipedia, verbatim
**7, 25**: Englische Wikipedia, übersetzt und angepasst
**26**: eigene Definition (auf Basis der deutschen Wikipedia)

/fh/// Fachhochschule St. Pölten

Compass-Verlag
Information für Ihren Vorteil

FH | JOANNEUM
University of Applied Sciences

Monopol

ÖGB VERLAG

SEMANTIC WEB COMPANY
linking data to knowledge

## *Über NOLDE*

Viele Verlagsunternehmen stehen heute vor der Situation, dass in den vergangenen Jahren große Mengen an unstrukturiertem Content (Freitext) angefallen sind, die nicht weiter verarbeitet und kommerzialisiert werden können. Gründe dafür sind fehlende Technologien und die dazugehörige Expertise. Das Projekt NOLDE – Network of Linked Data Excellence führt zwei Verlagsunternehmen (ÖGB Verlag, Monopol Verlag), zwei Technologieunternehmen (Semantic Web Company, Compass Verlag) und zwei akademische Partner (FH St. Pölten - Institut für Medienwirtschaft, FH JOANNEUM - Institut für Journalismus und Public Relations) zusammen, um neue Methoden des Enterprise Data Managements im Verlagswesen zu erproben. Im Kern steht die kommerzielle Erschließung von Linked Data Technologien für die Verbesserung von Produktions-Workflows und den Ausbau der bestehenden Service-Palette. Dazu gehören die Entwicklung neuartiger Content-Curation Services, Knowledge Discovery & Recommendation Services und Dynamic Semantic Publishing Strategien für die automatische Verarbeitung deutschsprachiger Contents.

**Aknowledgements**: Die Broschüre entstand im Rahmen des Projektes NOLDE – Network of Linked Data Excellence, welches von der Österreichischen Forschungsförderungsgesellschaft im Zeitraum September 2013 – Dezember 2015 unter der Projektnummer 841049 gefördert wurde. Die Leitung des NOLDE-Projektes oblag der Fachhochschule St. Pölten.

Kontakt: tassilo.pellegrini@fhstp.ac.at

http://nolde.fh-joanneum.at