Linked Data

Stefan Gradmann | stefan.gradmann@ibi.hu-berlin.de
Steffen Hennicke | steffen.hennicke@ibi.hu-berlin.de
Marlies Olensky | marlies.olensky@ibi.hu-berlin.de

Der Artikel beschreibt zunächst den Grundgedanken des Semantic Webs und verortet darin Linked Data. Danach werden die technischen Grundlagen von Linked Data anhand der vier „best practice"-Prinzipien von Tim-Berners Lee ausgeführt. Am Beispiel der Linked Open Data Cloud werden schließlich die Vor- und Nachteile von Linked Data erklärt und auf aktuelle und zukünftige Herausforderungen eingegangen.

Das Semantic Web und Linked Data

Das heutige Web enthält eine unüberschaubare Menge an Informationen und eine der größten Herausforderungen stellt hierbei die Qualität der gefundenen Informationen dar.

Ein einfaches Beispiel kann dies verdeutlichen: Eine Suche nach „Paris" in Google liefert 1.630.000.000 Ergebnisse, wobei die Suchmaschine nicht weiß, ob der Nutzer nach der französischen Hauptstadt „Paris", nach einer der amerikanischen Kleinstädte mit Namen „Paris", nach der mythischen Figur „Paris" oder nach einem Roman mit dem Titel „Paris" gesucht hat. Suchmaschinen können dem Nutzer hier nicht weiterhelfen, weil sie die Semantik und den Kontext der Dokumente des Webs, der sogenannten Ressourcen, nicht verstehen. Informationen und Dokumente im Web sind von Menschen für Menschen hinterlegt worden und diese Informationen kann die Maschine nicht differenziert verarbeiten. Nur Menschen verstehen den impliziten Kontext dieser Informationen und Dokumente. Wenn Maschinen über diese Kontextinformationen zu vorhandenen Ressourcen verfügen und diese somit differenzierter verarbeiten, können sie dem Nutzer auch besser helfen.

Das Semantic Web soll hier Abhilfe schaffen. Die Idee ist, das vorhandene Wissen im Web maschinenlesbar zu machen, indem Informationsressourcen kontextualisiert werden. Dazu muss dieser Kontext expliziert werden, was durch die Anreicherung mit Metadaten und die Verknüpfung mit sogenannten „Ontologien" erreicht wird. Das Versprechen lautet: "Just as the World Wide Web has revolutionized the way we connect and consume documents so can it revolutionize the way we discover, access, integrate and use data."[1] Nach dem Erfinder Tim Berners-Lee ist das Semantic Web eine Erweiterung des herkömmlichen Webs, in der Informationen mit eindeutigen Bedeutungen versehen werden, um die Arbeit zwischen Menschen und Maschinen zu erleichtern: Dabei sieht Tim Berners-Lee das Semantic Web als eine Erweiterung des herkömmlichen Webs, wo die Bedeutung von Informationen eindeutig definiert ist und dadurch die angesprochene Zusammenarbeit von Mensch und Maschine ermöglicht wird: „The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."[2]

Die Entwicklung des Semantic Webs wurde lange durch Konzepte der Künstlichen Intelligenz (KI) geprägt. "For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. Artificial-intelligence researchers have studied such systems since long before the Web was developed. (...) The challenge of the Semantic Web, therefore, is to provide

1 Heath, Bizer (2011): Linked Data: Evolving the Web into a Global Data Space, S. 2.

2 Berners-Lee, Tim; Hendler, James; Lassila, Ora (2001): The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In: Scientific American Magazine (May). Online verfügbar unter http://www.scientificamerican.com/article.cfm?id=the-semantic-web

a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web."[1] Dieser Ansatz geriet 2006/2007 in die Krise. Die damals gefundene Einsicht lautete, dass das Semantic Web vom Inhalt her aufgebaut werden muss, statt den Fokus zu stark auf die Struktur zu legen. An dieser Stelle kommt nun die Idee von Linked Data ins Spiel.

Das World Wide Web Consortium (W3C) fasst die Rolle von Linked Data wie folgt zusammen: Um das Semantic Web bzw. das Web of Data zu realisieren, müssen die Daten im Web zunächst in einem gemeinsamen und zugänglichen Standard vorliegen. Insbesondere die Beziehung zwischen diesen Daten muss explizit gemacht werden. Genau diese Sammlung von verbundenen Daten wird als Linked Data bezeichnet. Damit steht Linked Data im Zentrum des Semantic Webs und bildet die Grundlage für die umfassende Integration von und Inferenz über Datenbeständen im Web.[2] Diese Einsicht fasst Berners-Lee treffend zusammen: "Linked data is essential to actually connect the semantic web."[3] Das Semantic Web ist dabei das angesteuerte Ziel und Linked Data das Mittel, um dieses Ziel zu erreichen.

Prinzipien und technische Grundlagen von Linked Data

Tim Berners-Lee stellte 2006 vier *best practice*-Prinzipien für Linked Data auf.[4] Diese vier Prinzipien und die zugrundeliegenden technischen Standards[5] sollen im Folgenden kurz vorgestellt werden. In diesem Zusammenhang sei darauf hingewiesen, dass es sich hierbei um eine Empfehlung der *best practices* im Bereich Linked Data handelt und nicht um eine verbindliche technische Vorgabe. Eine entsprechende Diskussion zu Begriffsinhalt und Definition von Semantic Web und Linked (Open) Data ist derzeit im Gange.[6]

Use URIs as names for things.

Ein *Uniform Resource Identifier* (URI) ist ein einheitlicher Bezeichner für sog. Ressourcen wie Webseiten, sonstige Dokumente oder Video- und Audiodateien, Webservices etc. Derartige Ressourcen sind im Fall des klassischen Webs online aufrufbar und werden als *information resources* bezeichnet.

Im Kontext von Linked Data wird dieses Prinzip auf Ressourcen außerhalb des Webs ausgedehnt. Derartige *non-information resources* umfassen dann real-existierende Objekte und abstrakte Konzepte. Somit können konkrete Menschen, Orte oder Gegenstände sowie abstrakte Beziehungen zwischen diesen oder ganz allgemein menschliche Ideen und Vorstellungen bezeichnet und im Web repräsentiert werden.

Use HTTP URIs so that people can look up those names.

Das HTTP-Protokoll ist der universale Zugriffsmechanismus des heutigen Webs über den Ressourcen wie beispielsweise HTML-Webseiten erreichbar gemacht werden. Im Kontext von Linked Data werden HTTP-URIs verwendet, um Ressourcen sowohl eindeutig zu identifizieren als auch dereferenzierbar zu machen, d. h. der URI soll immer auch zu einer Beschreibung der jeweiligen Ressource führen und ist sozusagen ein Standard-Zeiger auf solche Ressourcen.

Dabei ist festzuhalten, dass der HTTP-URI tatsächlich den Namen für das bezeichnete Ding darstellt und nicht nur einfach eine eindeutige Adresse für dieses Ding bzw. eine Beschreibung für dieses im WWW.[7] Linked Data berücksichtigt keine anderen URI-Protokolle.

When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).

Der Erfolg des heutigen Webs basiert unter anderem auf der Durchsetzung der *HyperText Markup Language* (HTML) als einheitlicher Beschreibungsstandard für Dokumente. Für die Beschreibung und Veröffentlichung von strukturierten Daten wird daher ebenso ein einheitlicher Standard gefordert.

Das *Resource Description Framework*[8] (RDF) ist ein einfaches Graphen-basiertes Datenmodell für die Modellierung von Aussagen über Ressourcen in der Form von Tripeln (Subjekt – Prädikat – Objekt). In der Subjektposition steht eine Ressource, die über eine semantische Beziehung (property) mit einer anderen Ressource oder einem Literal in der Objektposition verbunden ist. Ressourcen und properties werden durch URIs identifiziert. Ist das Objekt eine Ressource, kann diese das Subjekt in einem weiteren Triple sein, wodurch ein semantisches Netz entsteht oder, wie es Tim Berners-Lee bezeichnet hat, ein *giant global graph*.[9]

Das *Ressource Description Framework Schema*[10] (RDFS) erlaubt die Definition von Klassen, mit denen Ressourcen typisiert werden sowie die Festlegung von properties, also erlaubten Verbindungen zwischen Ressourcen. Weiterhin kann die Semantik von Klassen und Properties in einer Taxonomie von Sub-Klassen und Sub-Properties weiter verfeinert und einfache (logische) Restriktionen (beispielsweise Domain/Range oder Kardinalität) können für die Verwendung von Ressourcen und Properties definiert werden. Mit RDFS lässt sich also ein Vokabular für die Modellierung von Linked Data in Form einer Ontologie beschreiben.

Die Abfragesprache *SPARQL Protocol And RDF Query Language*[11] schließlich ist das Instrument, um komplexe Abfra-

1 Ebd.
2 W3C. http://www.w3.org/standards/semanticweb/data [30.07.2011].
3 Berners-Lee (2006): Design Issues. Linked Data. http://www.w3.org/DesignIssues/LinkedData.html [01.08.11].
4 Ebd.
5 Vgl. ausführlicher dazu auch: Heath, Bizer (2011).
6 Campbell, MacNeill (2010): The Semantic Web, Linked and Open Data.
7 Vgl. dazu Walsh (2006): Names and Addresses. http://norman.walsh.name/2006/07/25/namesAndAddresses [30.07.2011].
8 RDF Primer des W3C: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/ [06.08.2011].
9 Berners-Lee (2007): Giant global graph. http://dig.csail.mit.edu/breadcrumbs/node/215 [30.07.2011].
10 RDFS Primer des W3C: http://www.w3.org/TR/rdf-schema/ [06.08.2011].
11 SPARQL Primer des W3C: http://www.w3.org/TR/rdf-sparql-query/ [06.08.2011]. SPARQL ist ein rekursives Akronym.

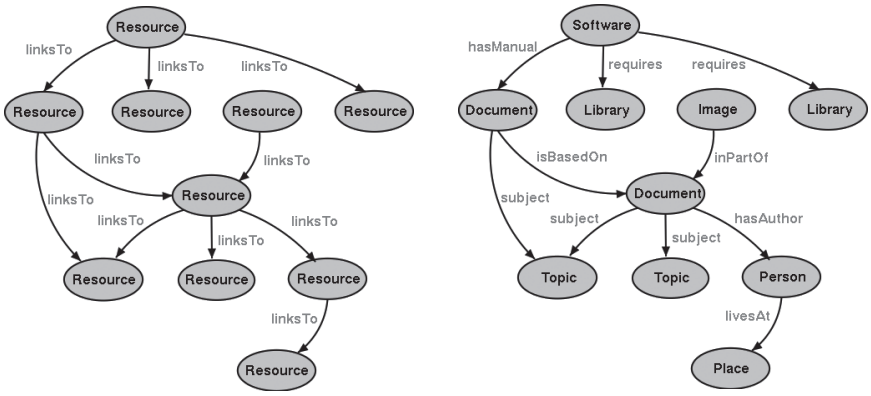

Abb. 1: Links die Struktur des herkömmlichen Webs und rechts die Struktur des Semantic Webs (beide Abbildungen aus Eric Millers Vortrag zum Semantic Web, zu finden unter http://www.w3.org/2004/Talks/0120-semweb-umich/).

gen über Datensätze im Linked Data Format zu formulieren. Zusammen mit RDF(S) bildet SPARQL daher die technische Grundlage für die Umsetzung des vierten Linked Data Prinzips.

Include links to other URIs, so that they can discover more things.

Das heutige Web besteht im Kern aus Ressourcen, vor allem aus HTML-Dokumenten, die über Hyperlinks miteinander verbunden sind. Diese Art der Verbindung kann nur von einem menschlichen Benutzer interpretiert werden.

Diese Links sollen, im Linked Data Kontext typisiert, die semantische Beziehung zwischen zwei Ressourcen also explizit und maschinenlesbar machen. Beispielsweise könnte ein sog. RDF-Link[1] die Beziehung zwischen zwei Ressourcen vom Typ Person als *isFriendOf* oder als *isColleagueof* charakterisieren oder die Beziehung zwischen einer Ressource vom Typ Person und einer Ressource vom Typ Ort mit *worksAt* beschreiben.

Die RDF-Links dienen also dazu, Ressourcen aus verschiedenen Datenquellen miteinander in semantischen Netzen mit maschinell prozessierbarer Syntax zu verbinden. Daten aus verschiedenen Wissensdomänen können somit in Verbindung treten und werden wiederverwendbar für verschiedene Anwendungszwecke. Durch die Verwendung gemeinsamer und offener Standards wie RDFS und SPARQL können Software-Applikationen implementiert werden, die über einen gemeinsamen Linked Data-Informationsraum operieren und komplexe Zusammenhänge zwischen Ressourcen entdecken und auswerten können.

Linked Open Data cloud

Dem Linked Open Data Projekt des W3Cs ist es zu verdanken, dass 2007 erste freie Datensets in RDF gemäß der Linked Data Prinzipien umgewandelt und im Web publiziert wurden.[2] Seitdem gilt das Prinzip, wer selbst Linked Data unter freier Lizenz publiziert, hat auch im Gegenzug Zugriff auf alle anderen Datensets in der sogenannten *Linked Open Data Cloud*.[3]

Die *Linked Open Data Cloud* bildet alle Datensätze ab, die im *Web of Data* im Linked Data Format veröffentlicht wurden. Ein entsprechendes Diagramm visualisiert die Zusammenhänge zwischen den einzelnen Datenquellen.

Die Größe der Kreise entspricht ungefähr der Anzahl der Triples in den jeweiligen Datensets (basierend auf den Schätzungen der Datenprovider), die Dicke der Pfeile gibt an, wie und wie viele RDF-Triples der jeweiligen Datensets untereinander verbunden sind. Die Farben der Kreise kennzeichnen die unterschiedlichen Themengebiete.[4] Themenübergreifende (*cross-domain*) Datensets sind besonders wichtig, um die einzelnen Themengebiete untereinander zu verbinden und eine Isolation der einzelnen *Web of Data*-Inseln zu verhindern. Im Zentrum dieser Datensets steht die DBpedia, die eine maschinenlesbare Version (in RDF) der Wikipedia ist: „The importance of DBPedia is not only that it includes Wikipedia data, but also that it incorporates links to *other* datasets on the Web, e.g., to Geonames. By providing those extra links (in terms of RDF triples) applications may exploit the extra (and possibly more precise) knowledge from other datasets when developing an application; by virtue of integrating facts from several datasets, the application may provide a much better user experience."[5]

Seit 2007 wächst die *LOD cloud* rapide. Die Erweiterung um neue Datensets bedeutet neue Verlinkungen und dies stellt wiederum ein gewaltiges Potential für neue Anwendungen im Bereich des Linked Data dar.

Vorteile und neue Möglichkeiten

Jede Organisation kann im Prinzip ihre Daten in der Linked Data Cloud publizieren, egal ob es sich hierbei um strukturierte (aus Datenbanken oder auch statische XML-Dateien) oder auch unstrukturierte Daten (Textdokumente) handelt. Für beide Fälle gibt es „Umwandlungsszenarien", die es relativ einfach ermöglichen, Linked Data zu produzieren und zu publizieren.[6] Hierzu besteht natürlich ein besonderer Vorteil von Linked Data in der Vernetzung der eigenen Daten und ihrer damit stattfindenden Kontextualisierung. Die eigenen Daten werden in ein Netzwerk von Daten eingebunden und dadurch insgesamt wertvoller. Zusätzlich trägt das Publizieren und Vernetzen von Linked Data dazu bei, Daten-Redundanz zu reduzieren – es gilt das Prinzip *reuse existing vocabularies*.

1 So bezeichnet zur Unterscheidung von herkömmlichen, nicht-typisierten Hyperlinks zwischen zwei Web-Dokumenten, vgl. Heath, Bizer (2011).

2 Heath, Bizer (2011), S. 30.

3 http://richard.cyganiak.de/2007/10/lod/ [06.08.2011].

4 Ebd.

5 W3C. http://www.w3.org/standards/semanticweb/data [30.07.2011].

6 Vgl. Heath, Bizer (2011), S. 70ff.

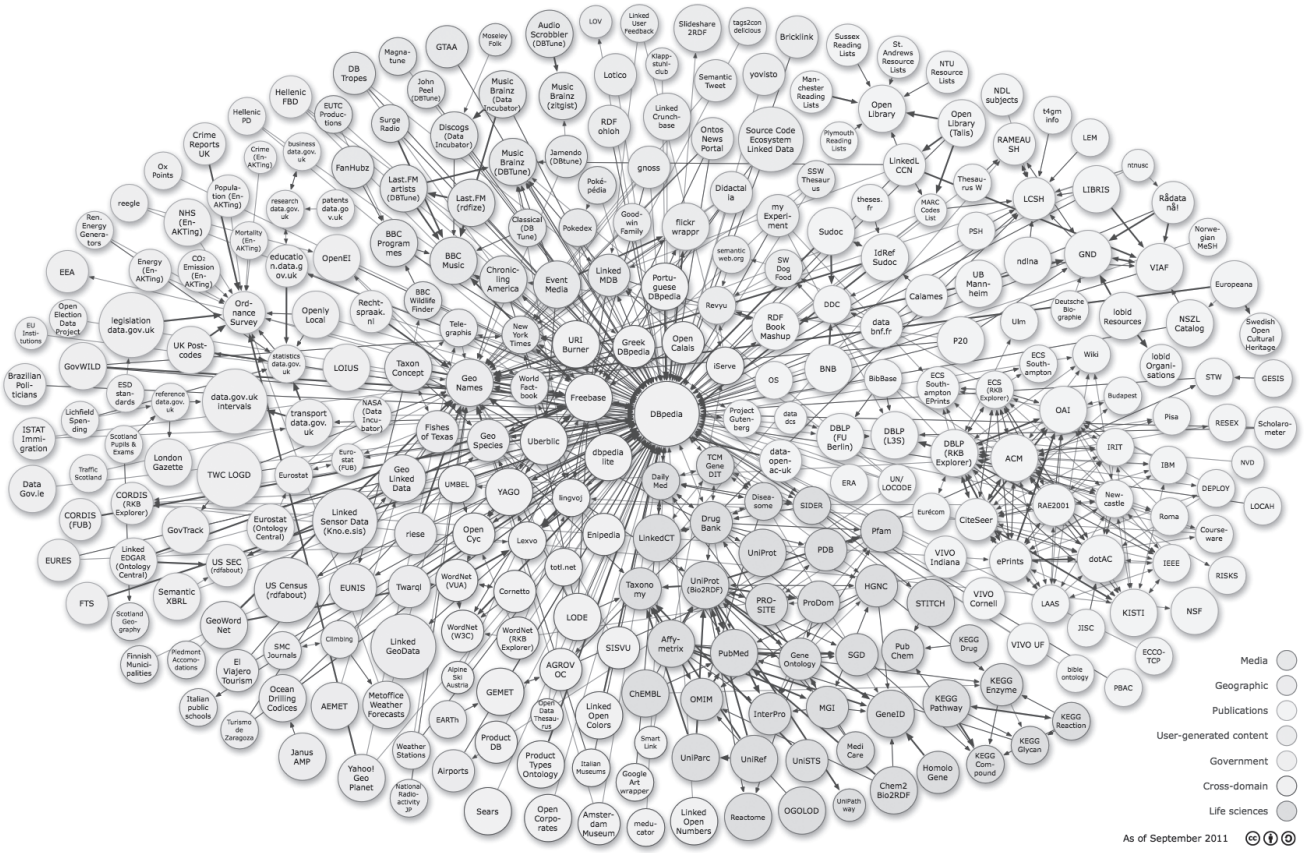

Abb. 2: Die Linked Open Data Cloud, Stand vom September 2011 (Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://richard.cyganiak.de/2007/10/lod/).

Umgekehrt ist es natürlich auch möglich, Linked Data zu konsumieren. Die Vorteile in der Nutzung von Linked Data bestehen klar in der standardisierten Datenrepräsentation und deren Zugang. Linked Data lässt sich somit viel leichter integrieren als proprietäre Web 2.0-Applikationen. Ein weiterer Vorteil ist die Offenheit des Web of Data, die ein Entdecken von neuen Datenquellen zur Laufzeit ermöglicht.[1] Linked Data-Browser und Linked Data-Suchmaschinen sind erste konkrete Anwendungen, die auch für den normalen Webuser die Vorteile von Linked Data sichtbar machen. Diese sind zwar meist noch als Prototypen im Testlauf, veranschaulichen aber den Mehrwert sehr eindeutig, z. B. Browser: Disco Hyperdata Browser (*http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/*), LinkSailor (*http://linksailor.com/nav*); Suchmaschinen: Sig.ma (*http://sig.ma/*), Swoogle (*http://swoogle.umbc.edu/*).

1 Heath, Bizer (2011), S. 85.

Schwierigkeiten und Baustellen

Aber genauso wie es im Web keine Kontrolle über den Inhalt publizierter Webseiten gibt, gibt es auch nur wenig bis gar keine Kontrolle im Web of Data. Dies soll natürlich dazu beitragen, dass möglichst viele Datensets publiziert und vernetzt werden, lässt aber somit zu, dass auch „schmutzige" Datensets ihren Weg in das Web of Data finden. Wie erwähnt, sind die Designprinzipien lediglich Empfehlungen, an die man sich halten sollte. Es gibt aber keine Instanz, die diese kontrolliert. Ein zusätzliches Problem, das ebenfalls schwer kontrollierbar ist, stellt das Wiederverwenden von bereits existierenden Vokabularien bzw. ‚Data Fusion' dar. „Data fusion is the process of integrating multiple data items representing the same real-world object into a single, consistent, and clean representation. The main challenge in data fusion is the resolution of data conflicts, i.e. choosing a value in situations where multiple sources provide different values for the same property of an object."[2] Ein Anspruch auf Vollständigkeit und Richtigkeit der Vernetzung der unterschiedlichen Datensets kann hier natürlich nicht erhoben werden. Auch die Herkunft der Daten in der Linked Data Cloud spielt hierbei eine Rolle. Die dereferenzierbaren URIs tragen dazu schon einen großen Teil bei, allerdings kann ein Namensraum auch von mehreren Datenprovidern verwendet werden. Um daher das Vertrauen in die publizierten Daten zu erhöhen, ist es hilfreich, auch Provenienz-Metadaten zur Verfügung zu stellen.[3]

Und schließlich wirkt sich im Web der Linked Open Data ein Designfehler schon des ursprünglichen Webs besonders stark aus: die Informationsarchitektur des Webs kennt keine Zeitbezüge, hat keine ‚Geschichte', ist faktisch nicht versionierbar. Das Web der Linked Data

2 Bizer, Heath, Berners-Lee (2009): Linked Data – The Story So Far.

3 Heath, Bizer (2011), S. 52.

kennt immer nur den zuletzt publizierten Zustand von Ressourcen und der Verbindungen unter ihnen. Welche Ressource vor beispielsweise einigen Monaten überhaupt schon vorhanden war und mit welchen anderen sie ggf. in welcher Art von Verbindung stand, ist mit den Mitteln des Webs nicht eruierbar. Abhilfe könnte hier nur eine systematische Reifikation der Graphen durch Meta-Aussagen schaffen (ein sehr umständlicher und wahrscheinlich wenig performanter Ansatz) – oder eine Erweiterung des RDF-Modells, das derzeit unter dem Rubrum *Named Graphs* diskutiert wird und Aussagen über Provenienz und Historie von RDF-Graphen möglich machen soll.

Schluss

Mit dem rapide wachsenden Web der Linked Open Data scheint die ‚Datenkrise' des Semantic Webs der ersten Generation gelöst, es entsteht nunmehr ein gigantischer, verteilter Informationsraum, in dem das Web auf Basis offener Standards wie HTTP und RDF zu einem „global data space"[1] wird.

Je mächtiger das damit gegebene Informationsuniversum wird und je mehr sich das Web damit zu einer gigantischen, erdumspannenden Maschine für die Wissensgenerierung und den Wissenstransfer wandelt, umso mehr verschiebt sich der Focus der Diskussionen in der ‚Community' weg von rein technikfixierten Auseinandersetzungen hin zu genuin politischen Fragestellungen – darunter die nach der intrinsischen Offenheit des Ansatzes. Die technischen Prinzipien der Linked Data erlauben gleichermaßen geschlossene, intranetbasierte Verfahren und das offene Netz der Linked Open Data, wenngleich erstere mit Restriktionen behaftet sind und niemals alle Möglichkeiten des offenen Ansatzes ausschöpfen werden, der ein unvergleichlich größeres heuristisches Potential aufweist. Es ist somit sehr wahrscheinlich (und wünschenswert!), dass die akademische Praxis im Sinne der Linked Open Data formiert sein wird – ob und wieweit dies auch im kommerziellen Anwendungsumfeld der Fall sein wird, ist derzeit nicht absehbar.

Ein wichtiges politisches Signal in diesem Sinne ist die wachsende Bereitschaft öffentlicher Einrichtungen vor allem in Großbritannien, aber zunehmend auch europaweit, die von ihnen mit Steuermitteln generierten Daten (Geodaten, Haushaltsdaten und viele andere mehr) als ‚Public Sector Information' (PSI) mit den Standards der Linked Open Data verfügbar zu machen!

1 Heath, Bizer (2011).

Literatur

[1] Berners-Lee, Tim: *Giant global graph*. 2007. http://dig.csail.mit.edu/breadcrumbs/node/215 [30.07.2011]

[2] Berners-Lee, Tim: *Design Issues*. Linked Data. 2006. http://www.w3.org/DesignIssues/LinkedData.html [01.08.2011]

[3] Berners-Lee, Tim; Hendler, James; Lassila, Ora: *The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*. Scientific American Magazine, May 2001. http://www.scientificamerican.com/article.cfm?id=the-semantic-web [03.08.2011]

[4] Bizer, Christian; Heath, Tom; Berners-Lee, Tim: *Linked Data. The Story so Far*. International Journal on Semantic Web and Information Systems, 2009. http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf [03.08.2011]

[5] Campbell, Lorna M.; MacNeill, Sheila: *The Semantic Web, Linked and Open Data*. A Briefing Paper. 2010. http://wiki.cetis.ac.uk/images/11/1a/The_Semantic_Web.pdf [03.08.2011]

[6] Cyganiak, Richard; Jentzsch, Anja: *Linking Open Data cloud diagram*. 2010. http://richard.cyganiak.de/2007/10/lod/ [12.08.2011]

[7] Heath, Tom; Bizer, Christian: *Linked Data: Evolving the Web into a Global Data Space*. 2011

[8] Miller, Eric: *The Semantic Web*. 2004. http://www.w3.org/2004/Talks/0120-semweb-umich/ [12.08.2011]

[9] Walsh: *Names and Addresses*. 2006. http://norman.walsh.name/2006/07/25/namesAndAddresses [30.07.2011]