Markus Luczak-Rösch, Ralf Heese und Adrian Paschke

Menschen und das Datenweb: Semantische Textverarbeitung für Nicht-Experten

Humans and the Web of Data: Semantic Text Authoring for the non-Expert

Textverarbeitung_Semantic Web_Linked Data_Annotation

Zusammenfassung. Durch die vielversprechende Linked Data Initiative ist das World Wide Web als ein Web das aus (HTML-) Dokumenten besteht, die mittels Hyperlinks vernetzt sind und dessen Informationen nur für die menschlichen Nutzer erschließbar sind, im Wandel zu einem so genannten Datenweb begriffen. Dennoch ist der Wert eines Datenwebs für Benutzer, die ohne technische Intention und Vorbildung das Web nutzen, bis heute kaum bekannt. Einige Merkmale semantischer Daten sind überdies nahezu unvereinbar mit dem etablierten Umgang mit Inhalten für das Web und der menschlichen Denk- und Arbeitsweise. Mit loomp präsentieren wir einen Ansatz für direkte semantische Textverarbeitung und die Verwaltung von Linked Data durch Nicht-Experten. Ziel ist es, der semantischen Annotation von Inhalten einen Vorrang vor der autorengetriebenen Formatierung zu geben und so den Zusatzaufwand, den manuelle Annotation bisher erzeugt, zu minimieren. Ferner wird eine Benutzerschnittstelle vorgestellt, die das Erzeugen, Verwalten und Durchsuchen von RDF-Daten im Datenweb in einer Form ermöglicht, die dem intuitiven Umgang der meisten technisch unerfahrenen Nutzer mit dem Web folgt.

Summary. Since the promising Linked Data initiative exists the World Wide Web as a Web of (HTML-) pages, which are interlaced by means of hyperlinks and whose information is capable of being developed for the human users only, is in the change to a so-called Web of Data. Nevertheless, the value of a Web of Data hardly admits for users, who use the Web without technical intention and training, until today. Some characteristics of semantic data are almost incompatible with established handling of Web contents and the human mindset and function. With loomp we present a prototype for direct semantic text processing and the administration of Linked Data by non-experts. A central goal is it to give the semantic annotation of contents a priority before the author-driven formatting and so to minimize the auxiliary expenditure, which is produced by manual annotation of new publications so far. Furthermore a user interface is presented, which facilitates the creation, administration and search of RDF on the Web of Data in a way, which follows the intuitive handling of the most technically unexperienced users.

1. Einleitung

Das World Wide Web als ein Web, das lediglich aus (HTML-)Dokumenten besteht, die mittels Hyperlinks vernetzt sind und dessen Informationen nur für die menschlichen Nutzer erschließbar sind, ist im Wandel begriffen. Die Erfolg versprechende Initiative zur Anreicherung von maschineninterpretierbaren RDF-Daten neben den konventionellen HTML-Inhalten heißt Linked Data – vernetzte Daten. Durch das gezielte Setzen von getypten RDF-Links zwischen RDF-Daten unterschiedlichster Quellen entsteht ein zweites Datenweb (Web of Data), das für die maschinelle Interpretation und Verarbeitung prädestiniert ist.

Schon heutzutage lassen sich interessante Anwendungsfälle und Geschäftsmodelle in der Literatur finden, die in Webseiten eingebettete RDF-Informationen nutzen. So entwickelte die BBC ein System, das die automatische Anreicherung des eigenen Inhalts mit nachgeladenen Informationen aus externen RDF-Datenquellen anwendet, um die Verweildauer von Besuchern auf den Webangeboten[1] zu erhöhen. Die Benutzer erhalten detailliertes Hintergrundwissen zu wichtigen Begriffen im Kontext der Website und müssen nicht eine externe Suche bemühen um sich dieses zu erschließen. Dennoch, im alltäglichen Gebrauch ist es weiterhin kompliziert den Geschäftswert und Nutzen zu ermitteln, den die Anreicherung und Veröffentlichung von semantischen Zusatzinformationen zu Inhalten erzielt. Das Argument einer verbesserten Suche auf Basis semantischer Zusatzinformationen allein ist nicht ausreichend, da diesem der zusätzliche Auf-

DOI 10.1524/icom.2009.0030

[1] http://www.bbc.co.uk/blogs/radiolabs/

wand zur Erzeugung semantischer Daten entgegen steht. Somit muss zwischen dem automatischen Erzeugen von semantischen Daten, die auf herkömmlichen Datenquellen basieren, und der manuellen Anreicherung unterschieden werden. Der erste Fall des Veröffentlichens von semantischen Daten auf automatische Weise, verursacht keinen zusätzlichen Aufwand für den Autor, bietet aber weniger bis keine individuelle Einflussnahme auf die Verteilung, Tiefe und Art der Anreicherung. Im zweiten Fall ist diese individuelle Einflussnahme für den Autor gegeben, sie erfordert jedoch mehr Aufwand für die Entwicklung des Inhalts.

Dieser Artikel präsentiert einen Ansatz für direkte semantische Textverarbeitung und die Verwaltung von Linked Data durch Nicht-Experten. Ziel ist es, der semantischen Anreicherung (Annotation) von Inhalten einen Vorrang vor der autorengetriebenen Formatierung zu geben und so den Zusatzaufwand, den manuelle Annotation bisher erzeugt, zu minimieren. Ferner wird eine Benutzerschnittstelle vorgestellt, die das Erzeugen, Verwalten und Durchsuchen von RDF-Daten im Datenweb in einer Form ermöglicht, die dem intuitiven Umgang der meisten technisch unerfahrenen Nutzer mit dem Web folgt. Damit soll die Lücke geschlossen werden, die zwischen maschinenverarbeitbarer Semantik von RDF und der menschlichen Denkweise und Erwartung im Bezug auf die Aggregation von Inhalten sowie deren Visualisierung besteht.

Wir haben den Artikel folgendermaßen strukturiert. Abschnitt 2 motiviert die Arbeit mit einem typischen Anwendungsfall aus dem Bereich des Journalismus. Im darauf folgenden Abschnitt werden Anforderungen an ein Werkzeug für semantische Textverarbeitung abgeleitet und im Abschnitt 4 die Architektur und Implementierung des entsprechenden Prototyps loomp[2] vorgestellt. Die Basis für den loomp[2]-Ansatz bildet die sogenannte Ein-Klick-Annotation (One-Click-Annotation), deren theoretisches Modell wir im fünften Abschnitt vorstellen und an loomp konkretisieren. Abschnitt 6 diskutiert die Nutzung und den Nutzen von semantisch angereicherten Daten im Rahmen der Visualisierung von Inhalten. Verwandten Arbeiten, eine Schlussfolgerung sowie der Ausblick auf weitere Arbeiten an diesem Thema schließen den Artikel in den Abschnitten 7 und 8 ab.

2. Analyse der Arbeitsweise im modernen Journalismus

In diesem Abschnitt stellen wir einen Anwendungsfall aus dem Bereich des modernen Journalismus vor. Dieses Szenario zeigt den Mehrwert, den das manuelle Erzeugen von Linked Data herbeiführen kann. Damit bildet es die Motivation für die Entwicklung eines Werkzeugs für Nicht-Experten zur direkten semantischen Textverarbeitung und Verwaltung von Linked Data. Dieser „Journalisten-Anwendungsfall" ist Repräsentant für alle Domänen mit inhalts- und wissensintensiver Arbeit in einem heterogenen Umfeld.

Unser Szenario beruht auf acht informellen persönlichen Interviews mit Journalisten sowie Redakteuren in Verlagen, die typischerweise in den Bereichen Print-, Online- und Cross-Media-Publishing arbeiten. Die Kernfragen dieser Interviews zielten darauf ab herauszufinden, wie der Workflow der Recherche und der Erstellung von Beiträgen aussieht, wie Beiträge angeliefert werden und wie Artikel endbearbeitet sowie kategorisiert werden, bevor sie auf unterschiedlichen Kanälen publiziert werden. Erstes Ergebnis nach der Befragung war, dass das Selbstverständnis der beiden oben genannten Zielgruppen nicht überschneidungsfrei darstellbar ist. So verstehen sich zum Beispiel Freiberufler, die Artikel auf Anfrage schreiben, selbst als Redaktueure. Für diesen Artikel vereinfachen wir dieses Verständnis und unterscheiden die beiden Gruppen anhand von Kernaufgaben: Journalisten, egal ob angestellt oder freiberuflich tätig, recherchieren und schreiben Artikel –- Redakteure revidieren und publizieren Artikel von Journalisten für einen Verlag oder Inhalteanbieter.

Im Kontext heutiger Verlage und Inhalteanbieter stellt sich dieses repräsentative Szenario von Journalisten und Redakteuren folgendermaßen dar. Journalisten recherchieren zu bestimmten Themen auf Anfrage und greifen auf verschiedene Informationsquellen zu diesem Zweck zu, z. B. Websites, Bücher, verwandte Artikel und menschliche Informanten. Unsere persönlichen Interviews ergaben, dass die Journalisten die Ergebnisse dieser Recherche unter Verwendung von Papier und Bleistift festhalten. Nur sehr wenige Journalisten verwenden digitale Medien für diese Aufgabe, noch weniger sogar Informations-Management-Systeme. Um den fertigen Artikel an den zuständigen Redakteur beim Verlag zu schicken, kommen häufig reine Text-Dokumente und E-Mail-Kommunikation zum Zuge. In einigen Fällen, vor allem im Online-Publishing-Bereich, sind Journalisten angehalten Artikel direkt in eine Redaktions-Management-System oder Content-Management-System einzugeben. Darüber hinaus sind dann entsprechende Kategorien und Schlagworte zum Artikel hinzuzufügen. Wichtig ist an dieser Stelle, dass die Formatierung der Texte an dieser Stelle kaum bis keine Relevanz hat, da das Setzen erst später in Abhängigkeit vom jeweiligen Publikationskanal und Medium vorgenommen wird. Schließlich prüft und revidiert ein Redakteur alle Beiträge im Ressort seiner Zuständigkeit und gibt eine Auswahl dieser abschließend zur Publikation frei.

In den letzten Jahren haben mehrere Studien wie die Delphistudie von Glotz (Glotz, 2009) aufgezeigt, dass Online-Publishing eine steigende Bedeutung hat und mehr professionelle Journalisten in

[2] http://www.loomp.org

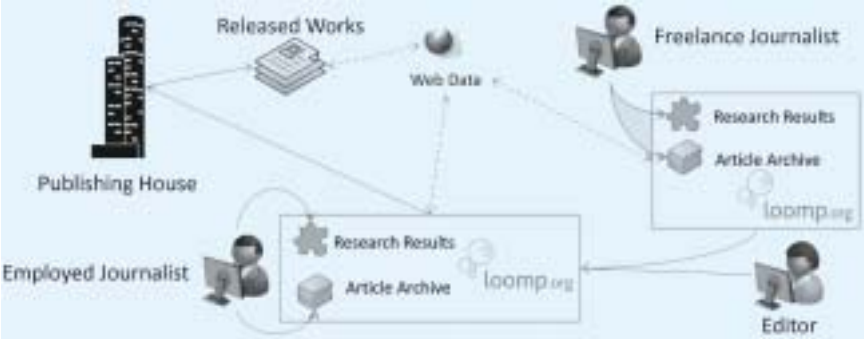

Bild 1: loomp im Kontext des Journalisten- Anwendungsfalls

Online-Medien gefordert sind. Klassische Verlage und Medienunternehmen sind die wichtigsten Informationsanbieter im Internet. In Bild 1 geben wir einen Überblick über den Journalisten-Anwendungsfall unter Verwendung eines modernen Content-Management-Systems (in der Darstellung bezeichnet mit dem Namen unserer prototypischen Entwicklung „loomp"), welches das Erstellen von textuellen Inhalten nebst semantischer Daten und deren Publikation als Linked Data in einem Arbeitsschritt zulässt. Die Anwendung dieses Content-Management-Systems ist dabei auf zwei Arten möglich:

- Wie ein Personal-Information-Management-System für die Journalisten, die von semantischer Suche und dem einfacheren Wiederverwenden von früheren Rechercheergebnissen und früheren Artikeln profitieren.
- Als Redaktions-Management-System für die Verlage, die das System als flexibles Instrument für Cross-Media-Publishing und personalisierte Content-Aggregation nutzen.

Das System hilft Journalisten ihre Notizen, Interview-Protokolle, Verweise (z. B. Hyperlinks) und Adressen zu verwalten. Der Anwender wird mit Hilfe eines intuitiven Editors dabei unterstützt, seine eingegebenen Textinhalte manuell semantisch zu annotieren. Das hilft insbesondere dabei einen Artikel und sämtliche relevante Informationsquellen zu verbinden. Die Darstellung der Inhalte ist sowohl für den Menschen lesbar als auch von Maschinen semantisch interpretierbar, so dass leichtes Finden, Wiederverwenden und Veröffentlichen möglich ist.

Auf Seiten der Verlage verwenden Redakteure diesen Ansatz zum Überarbeiten der Artikel, die von Journalisten für den Zugriff freigegeben sind. Sie können weitere Annotationen zu den Artikeln hinzufügen und möglicherweise Artikel unterschiedlicher Autoren miteinander verknüpfen. Schließlich wird ein Veröffentlichungs-Kanal für den Artikel gewählt (z. B. Blogs, RSS-Feeds, Wikis oder auch PDF-Dokumente als Vorstufe zum klassischen Printmedienum).

Die Vorteile dieser innovativen Infrastruktur sind vielfältig. Als wichtigstes Merkmal sei genannt, dass die semantische Suche von Inhalten möglich wird. Das bedeutet, dass nicht nur anhand von Zeichenketten im Volltext gesucht wird, sondern die Beziehungen der Annotationen sowie Dokument-Metadaten ausgenutzt werden, um relevant zusammenhängende Inhalte zu finden. Das bedeutet einerseits ein Verbreitern der Informationsmenge aufgrund automatischer Vernetzung von Inhalten, die gleichbedeutend annotiert sind. Im nächsten Schritt bedeutet es aber auch eine Schärfung der Suchergebnisse, da gleichlautende Begriffe mit unterschiedlicher Bedeutung abgegrenzt werden können. Da alle Inhalte angereichert werden, sind auch Szenarien möglich, in denen die Endnutzer die Präsentation von Inhalten nach dem aktuellen und individuellen Informationsbedarf verändern. Zum Beispiel kann der Leser entscheiden, alle Namen von Personen in einem Text zu markieren. Als Folge sind Journalisten und Redakteure deutlich von der Anstrengung des Formatierens von Texten in fett, kursiv oder unterstrichen befreit. Ferner können Inhalteanbieter, basierend auf den semantischen Annotationen, bessere zielgruppenspezifische Dienstleistungen anbieten.

3. Anforderungen an Werkzeuge für semantische Textverarbeitung

In diesem Abschnitt werden Anforderungen an ein Content-Management-System abgeleitet, welches die im vorhergehenden Abschnitt beschriebenen Charakteristika des motivierenden Anwendungsfalls erfüllt. Für das Design des Systems ist zu beachten, dass bei der Zielgruppe kein theoretisches Verständnis von RDF und den Linked Data Prinzipien vorauszusetzen ist. Darüber hinaus erwarten wir, dass sich das allgemeine Verständnis über das Web als ein Netzwerk von menschenlesbaren Seiten nicht in naher Zukunft ändern wird. Somit ist der Wert eines Datenwebs für Benutzer, die ohne technische Intention und Vorbildung das Web nutzen, bis heute kaum bekannt. Um die Nutzungsbarrieren unseres Systems zu senken, wollen wir die erfolgreichen Merkmale von Web 2.0-Anwendungen übertragen: leichtgewichtig, einfach zu verwenden und einfach zu verstehen. Wir beschreiben mit den folgenden Punkten Anforderungen an ein Linked-Data-Content-Management-System, die unserer Meinung nach notwendig sind, um auch unerfahrenen Benutzern die semantische Inhaltsanreicherung nahezubringen und sie als Autoren an einem Datenweb mitwirken zu lassen.

Intuitive Benutzeroberfläche

RDF-Informationen sind Aussagen über Ressourcen, wobei die Reihenfolge dieser Aussagen für die Maschinenverständlichkeit irrelevant ist. Dieses Paradigma widerspricht der menschlichen Denkweise in Artikeln, die eine gewollte Chronologie aufweist. Das System soll sämtliche RDF-eigene Erscheinungen bei der Erstellung von Linked Data für Autoren verbergen und eine intuitive Benutzeroberfläche bieten. Das Verfassen von Texten folgt der menschlichen Denkweise und nutzt bekannte Verfahren der System-Interaktion, um semantische Annotationen zu produzieren. Zum Beispiel ist jeder Computernutzer heute in der Lage, einen Text auszuwählen und diesen durch das Klicken auf eine Schaltfläche kursiv zu formatieren.

Einfache Vokabulare

Obwohl Web-Benutzer den Begriff URL oder Internet-Adresse kennen, ist ihnen derzeit nur selten bewusst, was Namensräume sind. Das System muss einen Zugang zu Annotationsvokabularen bieten, ohne dass der Nutzer die technischen Details der Repräsentation und Serialisierung dieser Vokabulare kennen muss. Jedes Konzept des Vokabulars hat eine sinnvolle Bezeichnung und ist in einfachen Worten erklärt. Das System unterstützt Standardvokabulare sowie mindestens weitgehend akzeptierte Vokabulare und ist in der Lage, Begriffe mit gleicher oder ähnlicher Bedeutung aufeinander abzubilden.

Assoziatives Verknüpfen von Inhalten

Grundsätzlich gehört die Verknüpfung von Inhalten über ein Dokument hinaus nicht zum Arbeitsablauf im Bereich klassischer Textverarbeitung. Das System muss auf Basis annotierter Begriffe und Phrasen Verknüpfungen zu anderen Inhalten automatisch herstellen.

Fokussierung auf die Aufgabe des Benutzers

In der Regel will ein Benutzer einen Text fortlaufend schreiben und nicht zusätz-

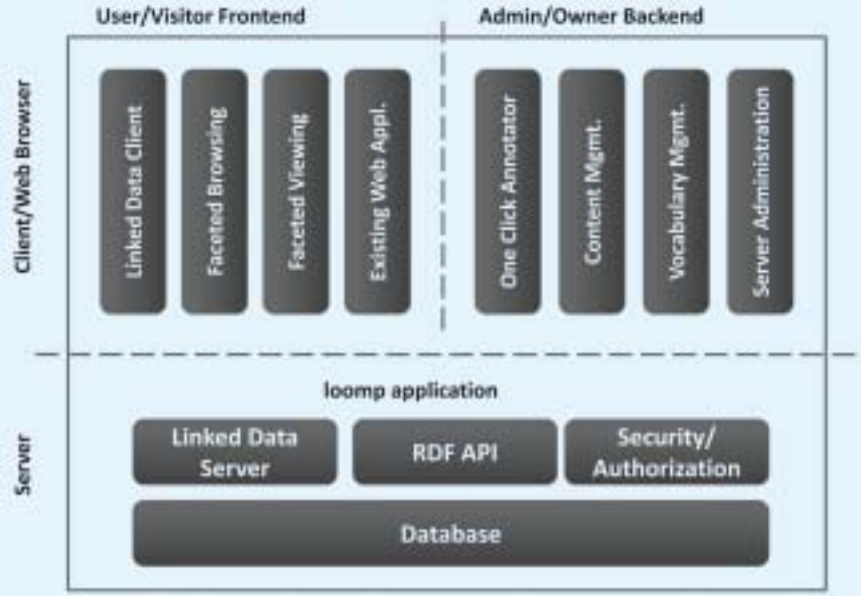

Bild 2: Einfache Übersicht der loomp-Systemarchitektur

liche Arbeitsschritte zur Anreicherung von Metainformationen im Nachgang durchführen. So muss die Symbolleiste für die Erstellung von semantischen Annotationen nahtlos in einem Text-Editor integriert sein, so dass der Benutzer Annotationen hinzufügen kann, ohne sich störend von seiner primären Aufgabe abwenden zu müssen.

Datenhoheit

Der Nutzer entscheidet selbst, welche Annotationen er einfügt, und kontrolliert, welche Daten veröffentlicht werden.

Einfache Installation

Die Anforderungen für die Installation und den Betrieb des Systems müssen gering sein, so dass es zum Beispiel in den meisten kostengünstigen Webspaces installiert werden kann. Die Notwendigkeit für die Konfiguration soll auf ein Minimum reduziert werden.

Einige dieser Anforderungen sind eher visionär, da neben technischen Anforderungen auch sozio-ökonomische Anforderungen erfüllt werden müssen. Beide, aber vor allem letztere, unterstreichen die Notwendigkeit und das Ziel, ein Content-Management-System für das Datenweb zu entwickeln, das nicht im Widerspruch zur menschlichen Mentalität und Denkweise funktioniert.

4. Architektur eines Werkzeugs zur semantischen Textverarbeitung

In diesem Abschnitt präsentieren wir die Anwendung mit dem Namen loomp als einen Prototyp für ein System zur semantischen Textverarbeitung und Linked-Data-Content-Management. Sowohl der Entwurf als auch die Implementierung dieses Werkzeugs folgen den von uns im vorherigen Abschnitt aufgelisteten Anforderungen.

loomp ist eine typische LAMP[3]-kompatible Web-Anwendung. Die grundsätzliche Architektur ist in Bild 2 dargestellt. Auf der Server-Seite sind die wichtigsten Komponenten eine Datenbank für die Speicherung sämtlicher Daten (Inhalte und Benutzerdaten), ein Linked-Data-Server für den Zugriff auf die RDF-Daten sowie eine RDF-API für den Zugriff auf und die Manipulation von RDF-Daten durch die Anwendung selbst. Selbstverständlich arbeitet im Hintergrund eine Komponente, die Sicherheit und Autorisierung für die Gewährung des Zugangs zu sämtlichen Daten regelt.

Auf der Clientseite wird zwischen Frontend und Backend unterschieden, wobei sich diese Begrifflichkeiten an deren Verwendung im Kontext klassischer Content-Management-Systeme orientieren. Der Begriff Frontend betrifft alle Bereiche, die mittels Daten- oder Web-Browsern ohne Authentifizierung zugreifbar sind. Die Komponenten Faceted Browsing und Faceted Viewing stellen Inhalte aus loomp in Webseitenform dar, wobei die als RDFa eingebettete semantische Annotation für die Navigation sowie die Änderung des Erscheinungsbildes genutzt werden kann. loomp verfügt über einen Plug-in-Mechanismus sowie eine API, die es ermöglicht von bestehenden Web-Applikationen aus (lesend und schreibend) auf den Inhalt zuzugreifen. So ist es zum Beispiel möglich, die annotierten Inhalte wie normale HTML-Seiten in einem CMS, Blog-System oder Wiki anzuzeigen. Das Backend umfasst Komponenten, für deren Zugriff eine Authentifizierung notwendig ist. Dies sind in der Regel die Komponenten zum Ändern und Verwalten der Inhalte (z. B. One-Click-Annotator und Content-Management).

In unserem System unterscheiden wir auf Datenebene zwischen zwei grundlegenden Arten von Ressourcen: Fragmente und Mashups. Ein Fragment stellt eine gedanklich geschlossene Einheit dar und besteht aus semantisch annotiertem Text oder einer SPARQL-Anfrage gegen eine externe RDF-Datenquelle. Mashups sind eine beliebige Anzahl von Fragmenten mit einer vom Nutzer festgelegten Reihenfolge. Sowohl Fragmente als auch Mashups sind durch einzigartige URIs identifiziert, die dazu dienen sie abzurufen oder ihnen Metadaten zuzuordnen. Im Einklang mit den Linked-Data-Prinzipien[4] kann ein Benutzer aus Fragmenten und Mashups Links zu anderen RDF-Datensätzen im Web (z. B. Konzepte innerhalb der DBpedia[5]) herstellen. Das System speichert sowohl eine XHTML/RDFa-Repräsentation der annotierten textuellen Inhalte, als auch eine reine RDF-Repräsentation aller extrahierten Annotationen nebst zusätzlicher Metadaten zum Autor, Datum, Datenschutz, etc. Für den Zugriff, die Manipulation und die Speicherung von RDF-Daten nutzen wir derzeit die RDF-API für PHP (RAP[6]). Die Gewinnung von RDFa wird von ARC[7] realisiert. Wir bedienen uns der RDF-Repräsentation für die Suche nach Ressourcen durch SPARQL-Anfragen sowie dem Apache Lucene Index[8] für die klassische Volltextsuche innerhalb der

[3] LAMP: Linux, Apache, MySQL, PHP/Perl
[4] http://www.w3.org/DesignIssues/LinkedData.html
[5] http://www.dbpedia.org
[6] http://www.seasr.org/wp-content/plugins/meandre/rdfapi-php/doc/
[7] http://arc.semsol.org/
[8] http://lucene.apache.org/

XHTML/RDFa-Repräsentation von FragmentenDurch den integrierten Linked-Data-Server bietet loomp auch direkten Zugriff auf die RDF-Daten aus einer Vielzahl von anderen Semantic-Web-Anwendungen heraus. Beispiele hierfür sind RDF-Daten-Browser (z. B. Disco, Tabulator, OpenLink Browser), RDF-Crawler (z. B. Syndice) oder in Anwendungen implementierte Anfrage-Bibliotheken (z. B. Semantic Web Client Library, SWIC). Der Linked-Data-Server bietet auch einen HTML-Template-basierten Mechanismus, um RDF-Daten in einer für Menschen lesbaren Form zu visualisieren.

5. Einfaches Annotieren und intuitives Nutzerinterface

Nachdem wir im letzten Abschnitt die Architektur von loomp, unseres Prototyps für semantische Textverwaltung und Linked-Data-Content-Management, vorgestellt haben, betrachten wir nun im Detail die zentrale Komponente des Systems: das innovative Nutzerinterface und den Mechanismus zur sogenannten „Ein-Klick-Annotation" (One-Click-Annotation, OCA).

5.1 Trennung von Inhalt, Bedeutung und Präsentation

Theoretisch setzen wir unseren Ansatz auf ein vom bekannten Paradigma zur Trennung von Inhalt und Präsentation/Layout abgeleiteten dreiteiligen Paradigma zur Trennung von Inhalt (Syntax), Bedeutung (Semantik) und Präsentation (Pragmatik) auf. Konkret ist dieses Konzept durch die Veränderung der Cascading Style Sheets (CSS) in Verbindung mit der XHTML/RDFa-Repräsentation der Inhalte realisiert, wie man in Bild 3 erkennen kann. Die Benutzer sind in der Lage die semantischen Annotationen zu nutzen, um kontextbezogen Ressourcen eines bestimmten Typs hervorzuheben. Dasselbe ist auch bezogen auf das Ausgabemedium oder den Publikationskanal in Form von Templates möglich.

5.2 Das Nutzerinterface zur Ein-Klick-Annotation

In (Heath et al., 2009) unterteilen die Autoren das Publizieren von Linked Data in die folgenden Schritte:

1. Auswählen passender Vokabulare
2. Unterteilen des RDF-Graphen in „Daten Seiten"
3. Zuweisen einer URI zu jeder „Daten-Seite"
4. Erstellen von HTML-Varianten der einzelnen „Daten-Seiten"
5. Zuweisen einer URI zu jeder Ressource auf diesen Datenseiten
6. Metadaten zur Seite hinzufügen und weitere Links auf andere RDF-Datenquellen setzen
7. Hinzufügen einer semantischen Sitemap[9]

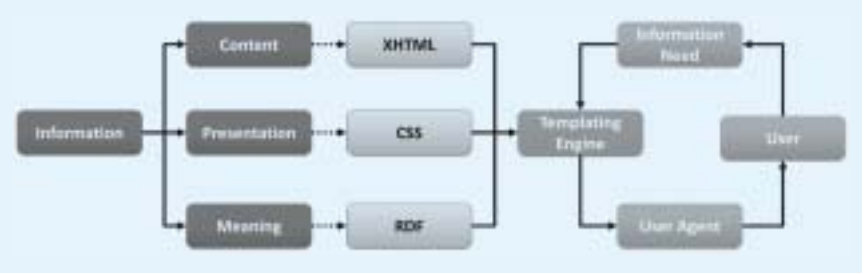

Bild 3: Darstellung der Trennung von Inhalt, Bedeutung und Darstellung in loomp

Um die Komplexität der Schritte 2 bis 6 zu reduzieren, die im Kern den Vorgang der Annotation von Inhalten beschreiben, orientiert sich die loomp Benutzerschnittstelle an bekannten Paradigmen moderner Textverarbeitungssysteme. Im Sinne der sogenannten Formatvorlagen haben wir eine Steuerleiste für unterschiedliche Gruppen zur semantischen Annotation implementiert. Jede Gruppe fasst thematisch zusammenhängende Annotationen zusammen. Beispiele hierfür wären die Annotationssätze für persönliche oder geographische Informationen, zwischen denen der Nutzer nach Bedarf Wechseln kann. Selbstverständlich können Annotationen unterschiedlicher Gruppen im selben Fragment gesetzt werden. Die Verwendung der Funktionsknöpfe der Steuerleiste erfolgt in bekannter Weise. Nach Auswählen eines Wortes oder einer Phrase im Text, indem dieser Teil markiert wird, kann durch Klicken die gewünschte Typisierung für die Auswahl festgelegt werden. Streng genommen kann eine Annotation auch dank dieser Benutzerschnittstelle praktisch nicht mit einem einzigen Klick realisiert werden. Dennoch nennen wir unseren Ansatz Ein-Klick-Annotation, weil der oben beschriebene fünfstufige Vorgang für den Nutzer auf das Vorhandensein einer einzigen Steuerleiste reduziert wird, die mit einem Klick mehrere der beschriebenen Aufgaben zusammengefasst im Hintergrund leistet.

Das Beispiel in Bild 4 zeigt die Annotation der Beziehung „firstname" für die Auswahl der Zeichenkette „Markus".

Im Hintergrund dieser Zuweisungsaktion durch den Benutzer wird auf Basis des Vokabulars, dass die Beziehung „firstname" definiert, bestimmt, welchen Typs die Ressource sein muss, die das System erzeugt. Die Vokabulare sind als RDF-Daten modelliert und stellen eine Teilmenge aus verschiedenen, von dritten gepflegten Standardvokabularen wie zum Beispiel FOAF oder Dublin Core dar. Sie sind um aussagekräftige Labels und Beschreibungen für die Anzeige in der Steuerleiste erweitert und definieren zusätzlich Abbildungen zwischen gleichbedeutenden Primitiven unterschiedlicher Vokabulare. Damit jede Ressource lediglich ein Mal im System über genau eine eindeutige URI erreichbar ist erscheint ein Dialog, der die Auswahl einer bestehen-

[9] http://sw.deri.org/2007/07/sitemapextension/

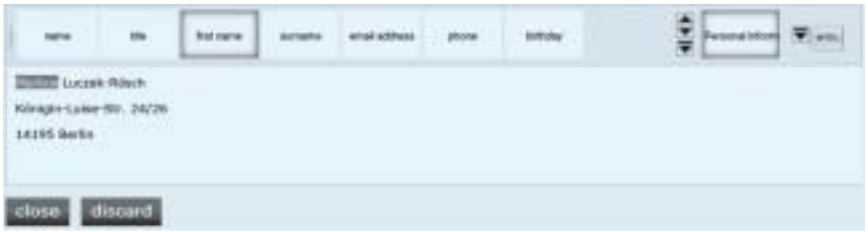

Bild 4: Benutzerschnittstelle zur Ein-Klick-Annotation in loomp

den oder das Anlegen einer neuen Ressource ermöglicht. Anschließend wird der gewählten oder angelegten Ressource als Metainformation die Beziehung „firstname" mit dem Wert der Zeichenkette „Markus" zugewiesen. Da es nicht das Ziel des globalen Datenwebs sein kann, dass in unterschiedlichen Datenquellen über dieselbe Entität mit unterschiedlichen URIs Aussagen getroffen werden, unterstützt loomp auch das Verwenden externer, global eindeutiger URIs. Hierfür ist das folgende Beispiel aussagekräftig. Die Zeichenkette „Berlin" soll in einem Text in der Beziehung „city" gekennzeichnet sein, wobei damit die deutsche Bundeshauptstadt gemeint ist. Selbstverständlich findet sich in großen Datensets wie DBpedia diese Ressource bereits mit einer eindeutigen URI. loomp sieht somit neben den Optionen eine vorhandene lokal eindeutig referenzierbare Ressource zu verwenden oder eine neue dieser Art zu erzeugen auch die Möglichkeit vor externe Datenquellen nach URIs für die ausgewählte Entität anzufragen. Wieder wird in Abhängigkeit vom Typ, der lokal zugewiesen werden soll, das Ergebnis der möglichen Ressourcen eingeschränkt.

Im Ergebnis hat der Nutzer mit der Ein-Klick-Annotation die Möglichkeit, während des Schreibens eines Artikels oder einer Notiz die Semantik wichtiger Wörter oder Phrasen manuell zu erfassen. Dabei wird unmittelbar auch die Option eröffnet eigene Inhalte durch Verwendung externer URIs für Ressourcen mit externen Inhalten zu vernetzen. Die Datenstruktur von RDF sowie technische Details der Serialisierung von RDF bleiben vor dem Nutzer verborgen.

5.3 Semantische Facetten und Visualisierung von Inhalten

Da der Nutzen von semantischer Textverarbeitung nur gegeben ist, wenn die Semantik auch für effizientere Informationsgewinnung genutzt werden kann, haben wir als ein Beispiel für eine innovative Visualisierungsmöglichkeit das facettierte Hervorheben von semantisch annotierten Wörtern oder Phrasen implementiert. Es realisiert die Anforderung, dass der Nutzer in Abhängigkeit von seinem Informationsbedürfnis bestimmte Bereiche von Inhalten hervorheben möchte, andere, die der Autor eventuell hervorgehoben hätte, hingegen nicht. Durch das Setzen von Filter-Facetten auf im Inhalt enthaltene Annotationen kann der Leser eines Textes bedeutend schneller die Bereiche identifizieren, die für ihn aktuell relevante Informationen enthalten.

Konkret wählt der Nutzer für das facettierte Hervorheben aus einer Liste von RDF-Primitiven, die im angezeigten Inhalt auch annotiert wurden und weist diesen Annotationen CSS-Stilinformationen wie zum Beispiel eine besondere Hintergrundfarbe zu. Umgehend wird diese Stilinformation auf alle im Inhalt identifizierten Bereiche, die mit dem gewählten Primitiv annotiert sind, angewandt. Es ist möglich auf diese Art sämtliche Annotationen in beliebigen Kombinationen hervorzuheben oder Hervorhebungen auch umgehend wieder aufzuheben.

Darüber hinaus ermöglicht unser System auch das Navigieren entlang der Semantik. Dies wird mittels eines Kontextmenüs realisiert, das mit einem rechten Mausklick für jede semantische Annotation in Inhalten verfügbar ist, wenn man sich in der Standardansicht eines Mashups, der XHTML/RDFa-Repräsentation, befindet. Das Menü bietet zum einen die Möglichkeit in den Modus des Daten-Browsings zu wechseln. Das heißt, beginnend an einer HTML-Visualisierung sämtlicher im System enthaltener Informationen der aktuell ausgewählten Ressource kann der Nutzer weitere damit in Verbindung stehende Ressourcen menschenlesbar visualisieren oder auch direkt als RDF/XML-Quelltext serialisieren. Die zweite Option des Kontextmenüs ist die Konzeptnavigation, die wiederum das Navigieren zu sämtlichen Mashups ermöglicht, welche die aktuell gewählte Ressource enthalten. Der Nutzer bewegt sich also weiter im Modus der XHTML/RDFa-Repräsentation von Mashups und navigiert anhand von gemeinsamen Ressourcen zwischen diesen.

6. Verwandte Arbeiten

Zu den bekanntesten Werkzeugen zum Erstellen von semantisch angereicherten Inhalten gehören semantische Wikisysteme. Beispiele hierfür sind OntoWiki (Auer et al., 2006), Ikewiki (Schaffert, 2006) oder Semantic MediaWiki (Krötzsch, Vrandecic, Völkel, 2006). Diese Wikisysteme erweitern Funktionalitäten von traditionellen Wikis um die Möglichkeit, mittels einer erweiterten Wikisyntax oder unter Zuhilfenahme von Inline-Editing-Modi, Annotationen zu Wiki-Seiten hinzufügen und getypte Beziehungen zwischen den Wiki-Seiten, basierend auf Ontologien, anzugeben. Unserer Meinung nach sind semantische Wikis noch weit davon entfernt von einer breiten Masse an Nicht-Experten nutzbar zu sein. Die vielmals notwendigen Bemühungen um eine spezielle Syntax zu lernen und die ursprüngliche Ausrichtung vieler dieser Systeme auf die Anzeige von Inhalten als Hypertext unterscheiden diese von loomp. loomp zielt grundsätzlich auf das kombinierte Veröffentlichen von Inhalten auf konventionellen digitalen Wegen (z. B. als PDF-Dokument), als Hypertext und im Datenweb ab.

Die Zemanta Semantic API[10] sowie OpenCalais[11] sind Beispiele für Dienste, die Inhalte automatisch mit Annotationen versehen. In loomp werden beide Dienste genutzt um dem Nutzer Vorschläge für Annotationen in Texten zu unterbreiten. Insgesamt stellt loomp eine Alternative gegenüber diesen Diensten dar, die den Nutzern die Möglichkeit bietet über die Art und Dichte der Annotationen in Abhängigkeit ihres Anwendungskontextes selbst zu entscheiden.

Auch automatische Systeme, die für Webseiten und relationale Datenbanken entwickelt wurden, um Linked Data zu generieren, sind im Kontext von loomp relevant. Die erwähnenswertesten Beispiele hierfür wären das Crawling der Wikipedia-Daten für das DBpedia-Projekt (Auer et al., 2007) sowie D2RQ-Server (Bizer und Seaborne, 2004), RDB2RDF (Malhotra, 2008) und Triplify (Auer et al., 2009) als Wrapper um relationale Datenbanktechnologien. Ähnlich wie im Verhältnis zu Zemanta und Open Calais grenzt sich loomp – und damit die Ein-Klick-Annotation generell – durch die gewollte Kontrolle des Nutzers über die tatsächlich eingefügten und veröffentlichten Annotationen ab.

7. Schlussfolgerungen und Ausblick

In diesem Artikel haben wir einen Ansatz für semantische Textverarbeitung vorge-

[10] http://www.zemanta.com/
[11] http://www.opencalais.com/

stellt. Wir haben den Zweck der semantischen Textverarbeitung im Rahmen der Problemstellung des Linked-Data-Content-Managements für den Nicht-Experten angesiedelt und durch einen Anwendungsfall aus der Domäne des Journalismus motiviert. Der Anwendungsfall basiert auf persönlich geführten, informellen Interviews mit Journalisten und Redakteuren und bildet die Grundlage für den Entwurf von Anforderungen, sowie die Basis für eine generelle Architektur für solche Werkzeuge. Ferner wurde das Nutzerinterface der prototypischen Implementierung, die den Namen loomp trägt, aus diesen Anforderungen abgeleitet.

Zusammenfassend kommen wir zu dem Schluss, dass es mit dem Prototyp loomp auch ungeübten Benutzern in kürzester Zeit möglich ist RDF-Daten zu erzeugen und zu verwalten. Nichtsdestotrotz fehlt zum heutigen Zeitpunkt noch eine echte Nutzerstudie, die nicht nur die Verwendbarkeit der loomp Applikation für Nicht-Experten fokussiert, sondern auch den generellen Vergleich zwischen klassischer und semantischer Textverarbeitung angeht. Wir sehen hierfür sowohl anwendungsabhängige, technische Kriterien als ausschlaggebend an (z. B. Precision und Recall bei der Inhaltssuche), vor allem aber auch qualitative Evaluationskriterien, die nur schwer darstellbar und auswertbar sind, weil sie unter anderem auch vom subjektiven Nutzerempfinden abhängen.

Danksagung

Dieser Artikel ist ein Ergebnis der Forschungsarbeiten im Vorhaben „InnoProfile: Corporate Semantic Web", das vom Bundesministerium für Bildung und Forschung (BMBF) und der BMBF-Innovationsinitiative für die Neuen Länder – Unternehmen Region – gefördert wird.

Neben den vielen Kollegen, die mit wertvollem Feedback die Arbeit an loomp befördert haben, danken die Autoren im Speziellen Tom Heath für seine Einladungsrede zur I-Semantics 2008 unter der Überschrift „Humans and the Web of Data", die inspirierend für die Entwicklung der Anwendung loomp und den Titel dieses Artikels gewirkt hat.

Literatur

Auer, S.; Dietzold, S.; Riechert, T.: OntoWiki – A Tool for Social, Semantic Collaboration. In: Cruz, I. F.; Decker, S.; Allemang, D.; Preist, C.; Schwabe, D.; Mika, P.; Uschold, M.; Aroyo, L., editors, *The Semantic Web – ISWC 2006, 5th International Semantic,* volume 4273 of Lecture Notes in Computer Science, Seiten 736–749. Springer, 2006.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z.: DBpedia: A nucleus for a web of open data. In: *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007),* volume 4825 of Lecture Notes in Computer Science, Seiten 715–728, Berlin, Heidelberg, Nov. 2007. Springer Verlag.

Auer, S.; Dietzold, S.; Lehmann, J.; Hellmann, S.; Aumueller, D.: Triplify - Light-weight Linked Data Publication from Relational Databases. In: *Proceedings of Semantic Data Web Track of 18th International World Wide Web Conference (WWW 2009),* April 20th–24th 2009, Madrid, Spain.

Bizer, C., Seaborne, A.: D2rq – treating non-rdf databases as virtual rdf graphs. In: *ISWC2004 (posters),* November 2004.

Glotz, P.; Meyer-Lucht, R.: *Zeitung und Zeitschrift in der digitalen Ökonomie – Delphi-Studie.* Project Website, Gesehen am 10. Januar 2009, http://www.unisg.ch/org/mcm/web.nsf/wwwPubInhalteGer/Online Publishing Delphi-Studie

Heath, T.; Hausenblas, M.; Bizer, C.; Cyganiak, R.; Hartig, O.: How to publish linked data on the web, Oktober 2008. *Tutorial at the 7th International Semantic Web Conference (ISWC2008),* gesehen am 10. Februar, 2009.

Krötzsch, M.; Vrandecic, D.; Völkel, M.: Semantic MediaWiki, Seiten 935–942. *Lecture Notes in Computer Science.* Springer Verlag, 2006.

Malhotra, A.: Progress report from the rdb2rdf xg. In: C. Bizer and A. Joshi, editors, *International Semantic Web Conference (Posters & Demos),* volume 401 of CEUR Workshop Proceedings. CEUR-WS.org, 2008.

Schaffert, S.: Ikewiki: A semantic wiki for collaborative knowledge management. In: *1st International Workshop on Semantic Technologies in Collaborative Applications (STICA'06),* Manchester, UK, June 2006.

1

2

3

1 Dipl.-Inform. Markus Luczak-Rösch ist wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Corporate Semantic Web an der Freien Universität Berlin. Sein Forschungsschwerpunkt liegt im Bereich der Verwaltung von Ontologielebenszyklen. Überdies ist Markus Luczak-Rösch als Gesellschafter der lumano – intelligent systems GbR im Bereich moderner Web-Technologien aktiv.
E-Mail: luczak@inf.fu-berlin.de

2 Ralf Heese ist wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Corporate Semantic Web an der Freien Universität Berlin. Zudem ist er seit Mai 2008 Gesellschafter der Ontonym -Gesellschaft für semantische Webanwendungen mbH.
E-Mail: rheese@informatik.hu-berlin.de

3 Prof. Dr. Adrian Paschke ist Direktor von RuleML Inc., gegründet in Kanada, Forschungsdirektor für R&D Partnerschaften am Center for Information Technology Transfer (CITT) GmbH Deutschland, und Professor für Corporate Semantic Web am Fachbereich für Informatik an der Freien Universität Berlin.
E-Mail: Paschke: paschke@inf.fu-berlin.de