Linked Open Data für die Exploration von Wissen im Web 2.0 mit SemaPlorer

Ansgar Scherp, Simon Schenk, Carsten Saathoff, Steffen Staab

Universität Koblenz-Landau

1 Hintergrund

Mit dem Übergang vom Web 1.0 zum Web 2.0 ist das Internet und sein Inhalt noch einmal spürbar gewachsen: Internetnutzer sehen sich einer kaum zu überschaubaren Flut an Informationen, Bildern und Applikationen gegenüber. Vor diesem Hintergrund erwächst auf Seiten von Nutzern aber auch kommerziellen Anbietern der Wunsch, Internetinhalte besser handhabbar zu machen. Ein Beispiel: Informationen über Reiseziele werden heutzutage selbstverständlich im Internet gesucht. Dazu werden zahlreiche Wikis, Portale und Webseiten aufgesucht, die eine unüberschaubare Anzahl von Texten, Bildern und Metainformationen enthalten, die von Internetnutzern online gestellt werden. Diese für den Benutzer schnell, sinnvoll und optisch ansprechend nutzbar zu machen, ist eine Herausforderung, der mit der von den Autoren entwickelten Java-basierten, Web 2.0 Anwendung SemaPlorer Rechnung getragen wird. Die SemaPlorer-Anwendung verknüpft verschiedene, sehr große Datenquellen unterschiedlicher Herkunft und Qualität auf intelligente Art und Weise und stellt sie dem Benutzer als sogenannter Web 2.0 Mashup dar. Eine zentrale Herausforderung ist dabei die Skalierbarkeit der sehr großen Datenmengen. Anstatt sich manuell durch Suchmaschinen und Portale begeben zu müssen, zeigen wir mit SemaPlorer eine Möglichkeit auf, die von Internetnutzern zur Verfügung gestellten Informationen sinnvoll und ihrem inhaltlichen Zusammenhang gemäß zu sortieren, anwenderbezogen und in einem angemessenen Zeitrahmen in einem einzigen System zu präsentieren und interaktiv erfahrbar zu machen. Die Datengrundlage von SemaPlorer bilden große semantische Datenbestände wie DBpedia, GeoNames, WordNet und persönliche FOAF-Dateien. Mit FOAF-Dateien können Benutzer ein Profil im Internet veröffentlichen, wie es mittels bekannter sozialer Netzwerkplattformen möglich ist. Über semantische Beschreibungen sind die einzelnen Datenbestände miteinander verknüpft. Sie sind außerdem verbunden mit einem großen Flickr-Datensatz, der mittels des Resource Description Frameworks (RDF)[1] semantisch beschrieben wird.

Ziel dieses Beitrages ist es, die wissenschaftlichen Hintergründe, die Entwicklung und den Nutzen der Java-basierten, Web 2.0 Mashup-Anwendung SemaPlorer für den Endanwender vorzustellen. Der Beitrag ist also sowohl aus wissenschaftlicher Sicht als auch aus Praxisgründen interessant und relevant. So wird mit der SemaPlorer-Anwendung eine der

[1] Resource Description Framework, http://www.w3.org/RDF/

führenden Technologien im Bereich der verteilten Anfrage von semantischen Datenquellen vorgestellt. Der Artikel demonstriert den Einsatz dieser Technologie in einem konkreten, praxisrelevanten Szenario und schafft damit den Transfer von der Forschung in die Praxis.

Die weitere Struktur des Artikels gliedert sich nun wie folgt: In Abschnitt 2 wird zunächst eine Motivation für kartenbasierte Anwendungen vorgenommen. In Abschnitt 3 wird die Java-basierte, Web 2.0-Anwendung SemaPlorer vorgestellt. Die für die SemaPlorer-Anwendung verwendeten Daten und deren semantische Verknüpfung werden in Abschnitt 4 eingeführt. Der Entwurf und die Architektur der SemaPlorer-Anwendung sowie die zu Grunde liegende Technologie zur verteilten semantischen Datenintegration werden in Abschnitt 5 präsentiert. In Abschnitt 6 wird eine formative Evaluierung der SemaPlorer-Anwendung vorgenommen. In Abschnitt 8 werden verwandte Arbeiten diskutiert und abschließend rundet eine Zusammenfassung den Artikel ab.

2 Motivation

Das Internet ist eine wichtige Quelle für Informationen über Städte, Urlaubsorte oder andere interessante Regionen. Heutige Anwendungen, die Nutzern für diese Aufgabe zur Verfügung stehen, sind zentralisiert und monolithisch, z.B. Reise-Websites wie Tripadvisor (http://www.tripadvisor.com) und Wikitravel (http://wikitravel.org) oder Wissensplattformen wie Freebase (http://www.freebase.com). Mit unserer neuartigen Infrastruktur und Web 2.0 Mashup-Anwendung SemaPlorer greifen wir auf ein Netz verbundener Datenbestände zu. Diese sind nahtlos in einer einzigen verteilten Infrastruktur integriert, um generischen Zugang zu den semantischen Multimedia-Daten zu erhalten. Die verschiedenen Datenbestände werden über SPARQL[2]-Endpunkte zur Verfügung gestellt. Über solche Endpunkte können semantische Datenbanken über die Anfragesprache SPARQL angesprochen werden. Damit können nahezu beliebige Datenquellen ad hoc zur Dateninfrastruktur von SemaPlorer hinzugefügt werden. Um Informationen aus dieser verteilten Infrastruktur abzurufen und zu visualisieren, bedienen wir uns mit der SemaPlorer-Anwendung dem sogenannten „Blended Browsing and Querying“-Ansatz (Munroe, Ludscher u. Papakanstantinou 2000). Die Nutzer können sich durch nahezu beliebige Datensätze unter Verwendung verschiedener Ansichten (Facetten) wie Ort, Zeit, Personen und Tags navigieren (Hearst 2006). Wenn

[2] SPARQL Query Language, http://www.w3.org/TR/rdf-sparql-query/

der Benutzer mit der Anwendung interagiert, werden dabei gleichzeitig mehrere Anfragen an die zugrunde liegende Speicher-Infrastruktur gesendet, um die entsprechenden Ergebnisse zu berechnen. Die Ergebnisse werden mittels einer Karte, Medien und verschiedenen Kontextansichten, die die verschiedenen Facetten repräsentieren, dargestellt.

Für SemaPlorer haben wir verschiedene semantische Datenquellen wie DBpedia (http://dbpedia.org), eine semantische Version von Wikipedia, GeoNames (http://geonames.org), eine umfangreiche Datenbank mit georeferenzierten Orten, WordNet (http://wordnet.princeton.edu) mit einer Abbildung des englischen Sprachvokabulars und persönliche FOAF-Dateien aus der semantischen Suchmaschine Swoogle (http://swoogle. umbc.edu) integriert. Darüber hinaus haben wir einen partiellen Crawl, also eine partielle lokale Kopie von Flickr (http://flickr.com) erstellt und als einen sehr großen, nicht-semantischen Datensatz, der umgerechnet auf 700 Mio. RDF Triple kommt, eingebunden. Der Datensatz umfasst alle Annotationen von Fotos auf Flickr von ca. Mai 2005 bis April 2006. Zusammen bilden diese Datenbestände einen sehr großen, semantisch heterogenen Datensatz von gemischter Qualität, die zusammen über eine Milliarde Triples ergeben. Die Verknüpfung dieser Daten erfordert eine flexible und skalierbare Speicherstruktur. Die SemaPlorer-Infrastruktur besteht aus 25 RDF-Datenbanken. Die Datenbanken werden in virtuellen Maschinen auf Amazons Elastic Computing Cloud (EC2, http://aws.amazon.com/ec2) gehostet. Die EC2 ist ein Dienst von Amazon, um eigene Anwendungen im Internet auszuführen und anzubieten. Der Simple Storage Service von Amazon (S3, http://aws.amazon. com/s3) wird genutzt, um die semantischen Datensätze zu speichern. Er stellt eine zu EC2 passende Infrastruktur zur Verfügung um große Datenmengen über das Internet bereitzustellen. Die Speicher können wie ein einziger, virtueller RDF Speicher über einen Federator angesprochen werden. Der Federator verwendet die von den Autoren entwickelte Technologie NetworkedGraphs (Schenk u. Staab 2008), einen SPARQL-basierten, verteilten View-Mechanismus für RDF und verteilte Auswertung von SPARQL-Anfragen (Schenk u. Petrak 2008; Zemaniek, Schenk u. Svatek 2008). NetworkedGraphs erlaubt einfaches, regelbasiertes Schließen zur Laufzeit, zum Beispiel für die Integration semantisch heterogener Daten. Das Verteilen von Anfragen innerhalb der Infrastruktur wird durch eine – ebenfalls RDF-basierte – Konfiguration gesteuert, die im Repository des Federators gespeichert ist. Diese Konfiguration kann zur Laufzeit des Systems angepasst werden. Daher wird das Hinzufügen neuer Datenquellen durch die Anpassung der Federator-Einstellungen extrem einfach, während sie für die SemaPlorer-Anwendung vollkommen transparent ist.

3 SemaPlorer-Anwendung

Die Suche nach Informationen über eine interessante Region, wie eine Stadt oder eine Ferienregion, ist eine Aufgabe, die oft über das Internet erledigt wird. Je komplexer diese Fragen sind, desto schwerer können heutzutage Suchmaschinen und Plattformen nützliche Informationen liefern. So lassen sich beispielweise Webseiten über Städte wie Berlin sehr einfach über Standard-Suchmaschinen wie Google finden. Andererseits ist es z.B. fast unmöglich, Orte mit Straßenkunst in Berlin zu finden. Diese Anfrage auf eine andere Stadt wie z.B. Paris zu übertragen, stellt eine zusätzliche Herausforderung für die die Anwendung dar, die die traditionellen Ansätze nicht lösen können. Mit der Java-basierten, Web 2.0 Mashup-Anwendung SemaPlorer unterstützen wir die Anwender bei der Durchführung solch komplexer Datenexplorationen über verschiedene Datenquellen hinweg. Dabei integrieren wir das Navigieren mit Hilfe von Facetten und die traditionelle Volltextsuche und erlauben dem Nutzer somit eine frühe Auflösung von möglicherweise mehrdeutigen Suchtermen. SemaPlorer unterstützt vier generische Facetten, nämlich Ort, Zeit, Personen und Tags. Andere Facetten können einfach konfiguriert und hinzugefügt werden.

Eine Facette kann verstanden werden als ein Filter für große Datenmengen. Zum Beispiel kann SemaPlorer die Sehenswürdigkeiten einer bestimmten Stadt oder Gegend unter der Verwendung der Ort-Facette filtern und darstellen und dabei ausschließlich Fotos von bestimmten Benutzern zeigen. Während der Benutzer mit SemaPlorer interagiert, werden unmittelbar verschiedene Anfragen im Hintergrund erstellt. Die Ergebnisse der Anfragen werden sofort in der visuellen Ansicht in der Anwendung hinzugefügt und dargestellt. Dieser Ansatz ermöglicht eine vom Nutzer gesteuerte Darstellung und interaktive Exploration der verwendeten semantischen Daten. In der SemaPlorer-Anwendung formuliert der Benutzer zunächst eine einfache Anfrage in Textform, die in der oberen linken Ecke von Abbildung 1 dargestellt ist. Die Ergebnisliste enthält verschiedene Orte, Personen und Tags, die der Anfrage entsprechen. Klickt der Benutzer beispielsweise auf die Stadt Berlin, aktualisiert SemaPlorer die Ansicht in der Mitte von Abbildung 1, welche eine Stadtkarte von Berlin zeigt. Gleichzeitig werden Anfragen ausgeführt und die Ergebnisse als Pins in der Karte dargestellt. Wiederum gleichzeitig werden Anfragen ausgeführt, die den rechten Teil von Abbildung 1 mit Kontextinformationen füllen.

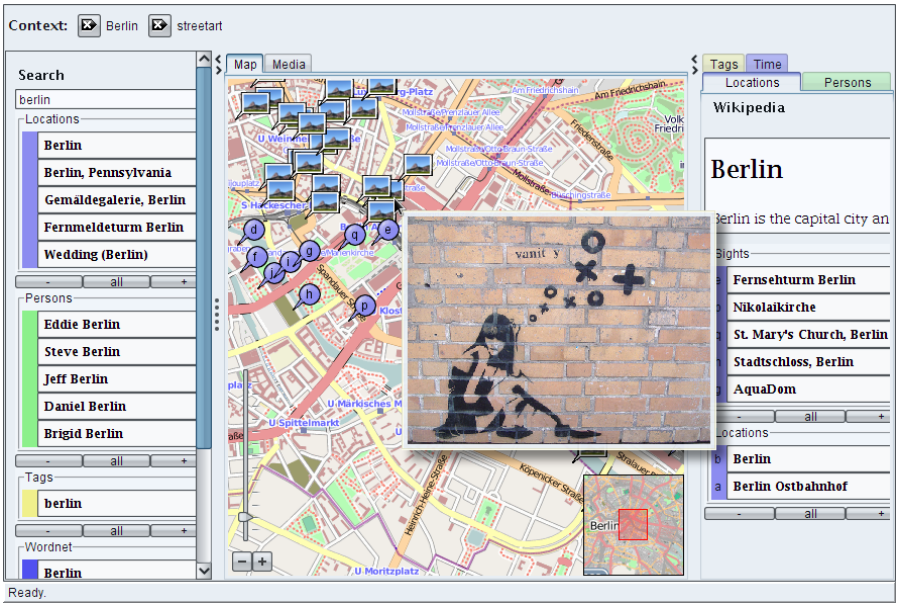

Abb. 1. Screenshot der SemaPlorer-Anwendung mit Straßenkunst in Berlin

Für jede Facette in SemaPlorer ist eine Kontextansicht definiert. Die Ort-Facette bietet z.B. Informationen aus DBpedia wie Bevölkerung und Land. Es werden Sehenswürdigkeiten und Orte in der Nähe gezeigt („nearby places“). Die Personen-Facette enthält Persönlichkeiten, die mit diesem Ort in Verbindung stehen, Flickr-Benutzer, die geo-referenzierte Bilder aus dieser Region hochgeladen haben und Internet-Nutzer, die in dieser Region leben – identifiziert anhand ihrer FOAF-Dateien. Die Zeit-Facette kann für die Auswahl eines speziellen Zeitraums wie beispielsweise Jahreszeiten wie Sommer und Winter genutzt werden. In der Tag-Facette werden Schlagworte von Flickr (Englisch: tags) dargestellt. Alle Facetten, wie Sehenswürdigkeiten, nahe gelegenen Orte, Persönlichkeiten und Tags, sind interaktiv. Dies bedeutet, dass die Benutzer über diese Facetten in den Kontextansichten navigieren können. Zum Beispiel können die Benutzer, wenn die Karte in SemaPlorer die Stadt Berlin zeigt, auf den Tag „street art“ (Straßenkunst) klicken. Sofort wird die Kartenansicht aktualisiert und die Standorte der Flickr Fotos welche mit „streetart“ annotiert sind, angezeigt. Durch die Eingabe einer Suche nach Paris kann der Nutzer zwischen dem aktuellen Kontext, nämlich Straßenkunst in Berlin, zu Straßenkunst in Paris wechseln und miteinander vergleichen.

4 Datensatz und Vernetzung der Daten

Um die facettierte, interaktive Suche und Visualisierung in SemaPlorer zu unterstützen, werden verschiedene Arten von semantischen Daten kombiniert. Wir verwenden einen signifikanten Teil der Datensätze, die für die Billion Triples Challenge[3] zur Verfügung gestellt worden sind, nämlich DBpedia (120 Millionen Triples), GeoNames (70 Millionen Triples), WordNet (2 Millionen Triples) und Swoogle (175 Millionen Triples). Darüber hinaus verwenden wir einen großen Datensatz von Flickr (700 Millionen Triples), der über mehrere Monate in den Jahren 2005-2006 gesammelt und in RDF übersetzt wurde. Wie in Abschnitt 2 beschrieben, haben wir verschiedene Kontextansichten für SemaPlorer definiert. Diese Kontextansichten ergeben sich aus den Eigenschaften beziehungsweise den zur Verfügung gestellten Informationen der verwendeten Daten. Im Folgenden beschreiben wir die verwendeten Daten entlang der vier in SemaPlorer definierten Facetten und wie sie miteinander verbunden sind.

Ort: Elemente dieser Facette beziehen sich auf die geographischen Koordinaten. Wir setzen GeoNames für Orte aller Art ein wie beispielsweise Städte, Länder und andere. Für Sehenswürdigkeiten verwenden wir eine Kombination der Volltext-Suche auf Artikelbeschreibungen aus DBpedia und deren Kategoriebeschreibungen, welche mit dem SKOS-Vokabular (Simple Knowledge Organization System, http://www.w3.org/2004/02/skos) beschrieben sind. Mit SKOS können Systeme zur Wissensorganisation beschrieben werden wie Thesauri, Klassifikationsschemas und Taxonomien. Zur Erkennung von Sehenswürdigkeiten betrachten wir die SKOS-Kategorien in DBpedia, insbesondere das Konzept SKOS:broader welches hierarchische Beziehungen zwischen SKOS Konzepten beschreibt und berechnen die transitive Hülle aller Kategorien. Außerdem nutzen wir eine Volltext-Suche auf den Katego-

[3] Ziel des Billion Triples Challenge (der internationalen Semantic Web Konferenz 2008 in Karlsruhe (http://challenge.semanticweb.org) war es, Anwendungen basierend auf großen semantischen Datensätzen zu entwickeln, die eine Echtzeitnutzung erlauben und dabei gleichzeitig einen deutlichen Mehrwert gegenüber bisher üblichen, einfachen Datenbankanfragen bieten. Der verwendete Datensatz musste dabei mindestens eine Größe von einer Milliarde (Englisch: 1 billion) atomaren Aussagen (Englisch: triples) im RDF-Format haben. Eine solche atomare Aussage beziehungsweise Triple besteht dabei aus drei Informationseinheiten, einem Subjekt, ein Prädikat und einem Objekt. So hat beispielsweise das Subjekt „Person“ das Prädikat „istGeborenIn“ mit dem Objekt „Stadt“ und sagt aus, dass Personen in einer Stadt geboren sind.

rienamen und schränken die Ergebnisse auf Einträge ein, die in der Kategorie dbpedia:Visitor_attractions einsortiert sind. Für die Anzeige der nahegelegenen Orte und Sehenswürdigkeiten wählen wir alle Geschwister eines gewählten Standortelements und sortieren sie auf der Grundlage der geografischen Distanz. Wenn z.B. der Arc de Triomphe in Paris ausgewählt wurde, werden als nahe Orte der Eiffelturm und Notre Dame angezeigt. Zusätzlich werden Bilder von Flickr dargestellt, die mit Geoinformationen versehen sind und sich im relevanten Kartenausschnitt befinden.

Zeit: Für die Zeit-Facette sind keine expliziten Daten definiert. Stattdessen können hier Inhalte aus einem bestimmten Zeitraum ausgewählt werden, z.B. Bilder aus einem bestimmten Monat aus Flickr. Darüber hinaus erlauben wir die Inhalte nach bestimmten Jahreszeiten wie im Sommer und Winter zu filtern.

Person: In den von SemaPlorer verwendeten Datensätzen haben wir drei Arten von Personen identifiziert: Diese sind Persönlichkeiten aus DBpedia, Flickr-Benutzer die Bilder eingestellt haben und Internet-Nutzer, die ihre FOAF-Dateien veröffentlicht haben und über Swoogle zugreifbar sind. Für jede dieser Kategorien von Personen verwenden wir eine andere Kombination der Daten. Für Persönlichkeiten wählen wir Bilder, die die ausgewählten Persönlichkeiten zeigen, basierend auf einer Volltext-Suche auf den Flickr-Tags. In Bezug auf einen Flickr-Nutzer suchen wir nach Inhalten, die durch den Benutzer veröffentlicht wurden. Für Internet-Nutzer betrachten wir den Geostandort in der FOAF-Datei (falls vorhanden) und verbinden sie mit Bildern von diesem Ort aus Flickr.

Tags: Tags stehen direkt im Zusammenhang mit den Flickr-Inhalten. Wir bieten Volltextsuche über die Tags. Wenn ein Tag von einem Nutzer ausgewählt wurde, zeigen wir verwandte Tags von Flickr sowie WordNet.

Komplexität der Anfragen: Um die oben beschriebenen Facetten mit Inhalten zu füllen werden mehrere Anfragen gleichzeitig ausgeführt. Für die initiale Suche mittels Stichworten, wie in Abschnitt 3 beschrieben, werden gleichzeitig drei Abfragen nach Orten, Personen und Tags durchgeführt. Bei einem Klick auf einen der gefundenen Einträge in der Ergebnisliste werden acht gleichzeitige Anfragen ausgeführt, um die Medien- und die Kartenansicht zu füllen, die nahegelegenen Orte zu berechnen, Sehenswürdigkeiten, Prominente und Flickr-Benutzer, Internet-Nutzer und Tags auszuwählen und die Zusammenfassung aus DBpedia zu erhalten. Die gleichen Anfragen werden durchgeführt, wenn der Kontext mit der aktuellen Ansicht geändert wird, z.B. wenn der Standort geändert wird, indem auf ein Bild oder einen Ort in der Nähe oder eine spezielle Person

geklickt oder ein Tag in der entsprechenden Kontextansicht gewählt wird. Dieser Ansatz ermöglicht eine sehr flexible Änderung der SemaPlorer-Anwendung um z. B. bestimmte Elemente der Ansichten hinzuzufügen oder zu entfernen. Die Anfragen nutzen die volle Ausdruckskraft von SPARQL. Darüber hinaus ermöglichen wir über die Textsuchmaschine Lucene bzw. LuceneSail eine Volltextsuche in SPARQL (http://dev.nepomuk.semanticdesktop.org/wiki/LuceneSail). Wir haben das LuceneSail erweitert, um Anfragen zu unterstützen, die die Form „A ODER B“ haben, und Anfragen nach geographischer Nähe zu ermöglichen. Die Anfragen bestehen typischerweise aus vier bis neun Joins, das heißt im Durchschnitt verbinden sie also bis zu vier Datensätze in einer einzigen Abfrage. Da die GeoNames- und Flickr-Datensätze über mehrere Repositories verteilt sind, werden für diese Datensätze jeweils mehrere Anfragen ausgeführt. Diese sind jedoch unkritisch, da sie leicht parallelisiert werden können. Je nach Kontext, den der Benutzer wählt, können die Anfragen komplexer werden, z. B. indem der Benutzer die Bilder auswählt, die mehrfach getagt sind und zusätzlich räumlich eingeschränkt sind.

Erfolge und Erfahrungen: Bei der Erstellung des Datensatzes für unsere SemaPlorer-Anwendung haben wir bemerkt, dass die Datensätze oft nicht vollständig und manchmal auch bezüglich der Semantik nicht eindeutig genug sind: Zum Beispiel fehlen in GeoNames zu einem beliebigen Eintrag Informationen über Sehenswürdigkeiten und Orte in der Nähe – Informationen, die in der HTML-Version vorhanden sind. Trotzdem konnten wir diese Informationen durch die Verbindung der einzelnen Datensätze, wie oben beschrieben, gewinnen. Des Weiteren haben wir beobachtet, dass die Daten auch innerhalb eines einzelnen Datensatzes heterogen sind. Zum Beispiel gibt es keinen klaren Lösungsansatz für die Angabe des Geburtsortes einer Person in DBpedia. Manchmal ist es dbpedia:cityofbirth und manchmal dbpedia:placeofbirth. In SemaPlorer lösen wir diese Unklarheiten durch die Zusammenfassung der beiden Eigenschaften in einem View. Ein View erlaubt eine bestimmte Sicht auf einen Datensatz zu legen und ermöglicht somit die Vereinheitlichung der beiden Modellierungsvarianten der DBpedia durchzuführen. Während Linked Open Data, also die Verknüpfung von semantischen Datenbeständen, fortschreitet, ist es immer noch eine offene Frage, wie es für die Verwaltung von Ressourcen wie Flickr-Bilder zu nutzen ist. Wie SemaPlorer zeigt, ist eine Kartierung der Linked Open Data und die semantische Beschreibung der Flickr-Daten in RDF möglich und funktioniert z.B. mit GeoNames gut. Doch statt der Kennzeichnung von Bildern mit Tags und anschließender Kartierung dieser Tags in Zusammenhang mit Open Data wäre es gewinnbringender, di-

rekt Linked Open Data zur Annotation, also Beschreibung der Bilder mittels semantischer Konzepte zu verwenden. Zum Beispiel könnte ein Bild, das den Eiffelturm zeigt, direkt mit dem entsprechenden Konzept für den Eifelturm aus der DBpedia annotiert werden.

5 SemaPlorer-Architektur

Die Architektur der SemaPlorer-Anwendung und Infrastruktur ist in Abbildung 2 dargestellt. Sie gliedert sich in zwei Sub-Systeme: Das erste Sub-System besteht aus dem K-Space Annotation Tool (KAT, https://launchpad.net/kat) und seinen SemaPlorer-spezifischen Erweiterungen, den KAT-Plugins. Es wird auf den Client-Computern eingesetzt und bietet die Benutzerschnittstelle und die Anwendungslogik von SemaPlorer. Das zweite Sub-System implementiert die verteilte Dateninfrastruktur und eine administrative Komponente für RDF-Repositories. Dies beinhaltet den NetworkGraphs-basierenden Federator und die verschiedenen RDF-Repositories für die semantischen Daten für die DBpedia-Abstracts und Flickr-Tags. Die Verwaltung der Komponenten und der Federator werden auf unserer EDV-Infrastruktur gehostet. Alle anderen Komponenten werden auf Amazon EC2 Knoten gehostet. Die Architektur von SemaPlorer und die einzelnen Komponenten werden im Detail im Folgenden beschrieben.

Das erste Subsystem, bestehend aus KAT und seinen Plugins, ist eine generische Architektur, die für die Entwicklung von Anwendungen für die Recherche und (semi-automatische) Annotation von Multimedia-Daten entwickelt worden ist. Es kann durch allgemeine Funktionalitäten wie einer interaktiven Karte oder den Zugriff auf Flickr-Bilder ergänzt werden. Ein Message-Bus erlaubt die Kommunikation der einzelnen Komponenten. KAT bietet einen Plugin-Manager für die Verwaltung anwendungsspezifischer Erweiterungen. Darüber hinaus bietet es einige GUI-Tools und einen GUI-Layouter. Schließlich verfügt KAT über eine lokale Speicherinfrastruktur für die Multimedia-Annotation auf Grundlage der COMM Multimedia Ontologie (Arndt et al. 2007) und SESAME 2 (http://openrdf.org).

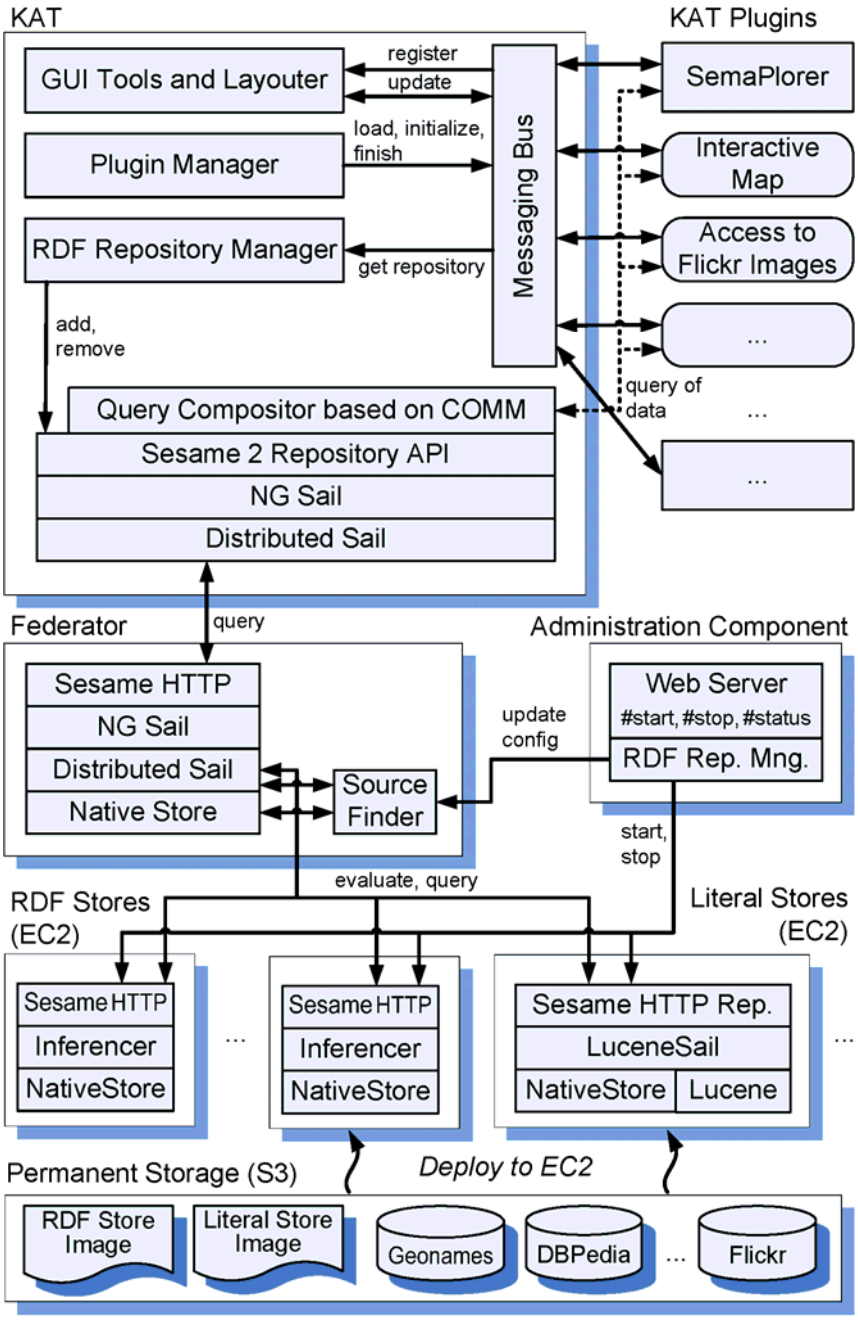

Abb. 2. Architektur von SemaPlorer

Der in Abschnitt 3 beschriebene Datensatz wird durch das zweite Subsystem, die auf NetworkedGraphs basierende verteilte Dateninfrastruktur unter Einsatz von Amazon EC2 realisiert. Die Verwaltungskomponente (Administration Component) dieser Dateninfrastruktur kontrolliert die virtuellen Maschinen, die auf EC2 laufen. Über eine einfach zu bedienende Web-GUI, können EC2-Knoten für spezielle Teile der Daten oder der gesamte Datenbestand gestartet und gestoppt werden. Neue Datensätze können durch Hinzufügen einer Beschreibung des Datensatzes zu einer Konfigurationsdatei und den Start des neuen Knotens geschaffen werden. Wenn Knoten gestartet oder gestoppt werden, aktualisiert die Verwaltungskomponente die Federator-Konfiguration. Der Federator ist der einzige SPARQL-Endpunkt, den SemaPlorer direkt nutzt, und verbirgt die Komplexität der unterliegenden Infrastruktur. Anfragen an den Federator werden analysiert, um festzustellen, welche Endpunkte zur Auswertung der Anfragen genutzt werden können. Anschließend wird die Anfrage in Unterabfragen aufgeteilt, die in den jeweiligen Repositories ausgewertet werden (Schenk, Petrak u. Sesame 2008; Zemanek, Schenk u. Svatek, 2008).

Die Datensätze werden dabei in Speicherknoten der EC2 mittels S3 gespeichert. Wir nutzen drei verschiedene Konfigurationen für EC2 Knoten: Die erste speichert RDF-Daten ohne jegliche Folgerungen. Sie wird z. B. für DBpedia Infobox Daten verwendet. Sie dient auch als Grundlage für die anderen beiden Knotentypen. Die zweite verwendet das LuceneSail und bietet zusätzlich Volltextindizes über die RDF-Literale. Es wird z. B. für die Tags, DBpedia Artikel- und Kategorienamen verwendet. Für die SemaPlorer-Anwendung brauchen wir kein vollständiges RDFS-Reasoning. Im Gegensatz dazu wird die Transitivität in SKOS-Hierarchien benötigt, die nicht in RDFS geboten wird. Daher verwenden wir benutzerdefinierte Regeln in der dritten Konfiguration von S3-Knoten. Da der entsprechende Inferenzer in Sesame nicht für den verwendeten Datensatz skaliert, führen wir eine Vorberechnung der transitiven Hülle von SKOS:broader für DBpedia-Kategorien durch. Vereinfacht gesagt berechnen wir also eine Erweiterung der SKOS:broader Relation, die zusätzlich alle indirekt erreichbaren Paare enthält (und damit transitiv ist). Der Federator erlaubt zudem die Definition von einfachen Views, um homogene Darstellungen aus den verschiedenen Datenquellen anzubieten. Ein Beispiel ist das oben genannte Vokabular für Geburtsorte in DBpedia. Dieses Schema-Mapping erfolgt während der Laufzeit mit NetworkedGraphs. Zum Beispiel haben wir für „Personen“ drei verschiedene Darstellungen: FOAF-Dateien, die das FOAF Vokabular-benutzen, DBpedia mit einer Personenkategorie und Flickr-Benutzer. Ähnliche Herausforderungen ergeben sich aus der Modellierung von räumlichen Einheiten und Annotatio-

nen von Bildern und bei Einträgen ohne ein klares Schema, wie der Geburtsort in DBpedia. Um SemaPlorer zu ermöglichen, von diesen verschiedenen Darstellungen zu abstrahieren, gestalten wir sie in einer kanonischen Form. Im Falle von Personen wird das FOAF-Vokabular verwendet. Als Ergebnis können wir jeden Datensatz-hinzufügen, der das FOAF-Vokabular einsetzt.

6 Evaluation

Ziel der Billion Triples Challenge ist es, die Skalierbarkeit von Semantischen Webtechnologien auf mehr als eine Milliarden Tripeln zu demonstrieren und damit etwas Sinnvolles zu tun. Als solches wurde die Java-basierte, Web 2.0 Mashup-Anwendung SemaPlorer und ihre zu Grunde liegende Infrastruktur als ein technischer Demonstrator, aber nicht als eine Endbenutzer-Anwendung, die in einer echten produktiven Umgebung läuft, entworfen. Um in einer solch frühen Phase eine Rückmeldung über die Benutzbarkeit und Nützlichkeit der Anwendung und Verbesserungsvorschläge zu erhalten, wurde eine formative Evaluation durchgeführt. Wir baten 20 Personen aus dem Institut für Informatik der Universität Koblenz-Landau (11 Doktoranden, 9 Studierende), SemaPlorer auszuprobieren. Die Personen sind zwischen 21 und 26 Jahren alt und haben gute bis sehr gute Computer-Kenntnisse. 18 Teilnehmer haben bereits Erfahrung mit der Nutzung von Karten-basierten Anwendungen zur Informationsbeschaffung und Visualisierung wie z. B. Google-Maps. Sie verwenden diese Anwendungen für Reiseplanungen (75%), um Informationen über den Ort zu erhalten (55%) und für berufliche Zwecke (25%). Demzufolge sind die Testkandidaten typische Benutzer von Anwendungen wie SemaPlorer und sind gute Kandidaten zur Ermittlung relevanter Rückmeldungen.

Die Evaluation der SemaPlorer-Anwendung wurde in drei Phasen unterteilt, nämlich Einführung, Test und Rückmeldung. In der Einführungsphase wurden die Teilnehmer mit SemaPlorer und seinen Features vertraut gemacht. Den Teilnehmern wurde erklärt, dass es nicht um die Messung ihrer Leistungen bei der Abarbeitung der Evaluationsaufgaben geht, sondern um die Gewinnung von Erkenntnissen zur Verbesserung von SemaPlorer. In der nachfolgenden Testphase erfolgte die eigentliche Bewertung. Jeder Teilnehmer hatte eine festgelegte Zahl von Aufgaben auszuführen. Eine solch einheitliche Aufgabenstellung ist wichtig, um eine Vergleichbarkeit zwischen den einzelnen Testpersonen herstellen zu können und eine valide Rückmeldung zu erhalten. In der Feedback-Phase füllten die

Teilnehmer einen Fragebogen aus. Die Fragen zur Erfassung der Zufriedenheit der Teilnehmer haben wir in Anlehnung an den IsoMetrics-L Fragebogen erstellt. Es wurde jedoch keine explizite Gewichtung der einzelnen Fragen vorgenommen, sondern den Benutzern die Möglichkeit gegeben, punktuell subjektive Rückmeldungen zu den einzelnen Fragen bzw. Funktionalitäten der Anwendung die sie für wichtig erachten zu geben. Für die Testphase hatten die Teilnehmer keine Zeitbegrenzung, sie konnten sich so viel Zeit lassen, wie sie benötigen, um die Aufgaben zu erfüllen. Die Dauer der Sitzungen lag zwischen 10 und 60 Minuten (Durchschnitt 30, Median 25). Demzufolge haben die Teilnehmer eine angemessene Zeit mit der Lösung der Aufgaben verbracht. Die Aufgaben, die ausgeführt werden sollten, sind die Suche nach der Stadt Berlin und die Verwendung des „Sights"-Features gewesen. Dann sollte das Ergebnis durch Hinzufügen des „streetart"-Tags auf die Anzeige von Bildern zu Straßenkunst eingeschränkt werden und Bilder mit Straßenkunst rund um den Berliner Sendeturm unter Benutzung des „nearby places"-Features erkundet werden. Eine spezielle Form der Straßenkunst sind „Space Invaders"-Piktogramme, die durch Hinzufügen des „spaceinvaders"-Tag gefunden werden. Die Teilnehmer sind gebeten worden, „Space Invaders" in Berlin zu finden. Anschließend sollte der Ortsbezug auf Paris abgeändert werden, um dort „Space Invaders" zu suchen. Um Paris weiter zu erkunden, wurden die Testpersonen gebeten, nach bestimmten Flickr-Usern und interessanten Bilder, die diese aufgenommen haben, zu suchen. Zusätzlich sollten die Testpersonen nach Persönlichkeiten in Paris in DBpedia suchen. Schließlich sollten die Benutzer entlang semantischer Relation zu dem Wort Paris in WordNet navigieren.

In der letzten Phase wurden die Testpersonen gebeten, einen Fragebogen auszufüllen, um damit Feedback über die bereits implementierten Features in SemaPlorer und die Anwendung als Ganzes zu bekommen. Tabelle 1 fasst die Fragen und die Beurteilungen zusammen. Die Fragen konnten gemäß IsoMetrics auf einer Skala von 1 bis 5 bewertet werden, bei der 1 „absolut keine Zustimmung", 2 „keine Zustimmung", 3 „teils-teils", 4 „Zustimmung", 5 „absolute Zustimmung" bedeutet. Die Rückmeldungen zeigen, dass die Suchergebnisse die Erwartungen der Nutzer erfüllen (S1). Ebenso wurde die Aufteilung der Suchergebnisse in Orte, Tags und Personen begrüßt (S2). Trotzdem könnte der Kontextwechsel durch die Nutzung der Suchfunktion intuitiver gestaltet werden (S3). Die Benutzbarkeit von Karten- und Medienansicht wurde als gut oder besser beurteilt (A1 und A2). Bei der Beurteilung der einzelnen Features der Facetten wurden alle Features als gut oder durchschnittlich bewertet, z.B. die Auswahl der Ansichten in der Ort-Facette (F1). Es wurden auch interessante Ansichten

gefunden (F2). Das „nearby places“-Feature wurde als gut bewertet (F3). Allerdings wurde die Qualität der gefundenen „nearby places“ nur mit teils-teils bewertet – hier sollten Verbesserungen erfolgen. Die Navigation entlang von WordNet und die Auswahl von Prominenten aus DBpedia (F6 und F7) wurden beide als teils-teils bewertet. Wir denken, dass hier insbesondere das Feature der Navigation entlang WordNet zu hinterfragen ist und eventuell entfernt werden sollte. Nur die Funktion, über Flickr-User zu navigieren, wurde von den Teilnehmern abgelehnt. Anscheinend lieferte diese Suchfunktion nur sehr wenige oder uninteressante Bilder von Persönlichkeiten oder Flickr-Benutzern (F8 und F9).

Tabelle 1. Feedback zur Suchfunktion (S1-S3), Karten- und Medienansicht (A1-A2), sowie den Facetten (F1-F9) und der Performance (P1) der SemaPlorer-Anwendung

Frage	*Mittelwert*	*Standard-abweichung*
S1: Die Suchergebnisse entsprechen meinen Erwartungen.	3.3	0.9
S2: Die Aufteilung in Orte, Tags und Personen ist intuitiv.	2.8	0.7
S3: Der Kontextwechsel mittels der Suchfunktion ist intuitiv.	1.8	1.0
A1: Die Kartenansicht ist intuitiv und einfach zu benutzen.	3.0	0.6
A2: Die Medienansicht ist eine gute Ergänzung zur Kartenansicht.	3.2	0.8
F1: Ist die Funktion zur Auswahl von Sehenswürdigkeiten sinnvoll?	3.4	0.5
F2: Haben Sie interessante Sehenswürdigkeiten gefunden?	2.8	0.7
F3: Ist die Funktion “nearby places” sinnvoll?	3.1	0.6
F4: Haben Sie interessante “nearby places” gefunden?	2.2	0.9
F5: Ist die Navigation mittels WordNet sinnvoll?	2.1	1.0
F6: Haben Sie interessante Persönlichkeiten in DBpedia gefunden?	2.4	1.0
F7: Ist diese Funktion sinnvoll?	2.4	1.0
F8: Haben Sie interessante Flickr-Benutzer gefunden?	0.9	0.8
F9: Ist diese Funktion sinnvoll?	1.7	1.0
P1: Die Anwortzeiten der Anwendung entsprechen meinen Erwartungen.	2.5	1.2

In der letzten Phase unserer Evaluation konnten die Testpersonen Feedback zu den in den Fragen genannten Funktionen geben sowie Vorschläge für weitere Funktionen machen, die sie gerne in SemaPlorer hinzufügen würden. So wurden grundsätzlich alle existierenden Funktionen zur Suche, Karten- und Medienansicht und den Facetten begrüßt. Lediglich das Browsen über WordNet, die Suche nach Persönlichkeiten in DBpedia sowie nach Flickr-Benutzern wurde von vielen Testpersonen als nicht sinnvoll erachtet, da keine passenden Ergebnisse gefunden werden konnten.

Fünf von 20 Personen schrieben, dass sie eine Erhöhung der Performanz von SemaPlorer begrüßen würden. Obwohl die Antwortzeiten im Allgemeinen gut waren, so haben komplexere Anfragen mehr Zeit gebraucht, als sich die Tester wünschten. Im Fragebogenteil wurde die Antwortzeit von gut bis teils-teils bewertet (P1). Diese Einstufung mag zunächst überraschen, aber wir gehen davon aus, dass den Testpersonen kommerzielle Produkte wie Google Maps als Vergleich dienten. Daher ist es wichtig zu betonen, dass SemaPlorer keine Anwendung ist, die auf einem Produktiv-Server läuft wie z.B. Google Maps, sondern eine technische Demonstration ist, die die Skalierbarkeit von Semantic Web Technologien zeigt. Außerdem wurden einige Vorschläge für Verbesserungen zur Benutzbarkeit der Anwendung gemacht, wie zum Beispiel den Ortswechsel über das Facetten-Menü intuitiver zu gestalten.

Hinsichtlich zusätzlicher Funktionalitäten wurde zum Beispiel eine Verlaufsfunktion genannt, welche das Vor- und Zurückspringen in den Navigationsschritten ermöglicht, die Auswahl mehrerer Orte, um eine Reise zu planen und die Präsentation einer Slideshow der Bilder. Eine Person fügte als Anmerkung hinzu, dass bereits zu viele Features vorhanden sind.

Wir fragten die Testpersonen außerdem, welche zusätzlichen Datenquellen wir SemaPlorer hinzufügen sollten. Hier wurden unter anderem die Integration von Satellitenbildern, weitere Medientypen wie Video, Nachrichten, andere Ansichten wie U-Bahn-Stationen, Cafés und Kinos sowie Meta-Informationen über Sehenswürdigkeiten wie Öffnungszeiten genannt. Sehr interessant war der Vorschlag, ein Bewertungssystem für die Vertrauenswürdigkeit der gelieferten Informationen einzubauen.

7 Verwandte Arbeiten

Der Grundgedanke der facettierten, interaktiven Suche ist die Exploration von großen Datenmengen und ist seit längerem bekannt (Yee et al. 2003).

Der Gewinner der Semantic Web Challenge 2006, /facet (Schraefel et al. 2005), hat diese Idee in den Bereich von semantischen Daten eingebracht. Vor kurzem ist die Anwendung Freebase Parallax (http://mqlx.com/~david/parallax) veröffentlicht worden, ein facettierter Browser für Exploration und Visualisierung der strukturierten Daten von Freebase (http://www.freebase.com). Der größte Nachteil von /facet und Freebase Parallax ist, dass sie auf zentralisierten Infrastrukturen basieren, die keinen skalierbaren Einsatz von einer großen Anzahl von Daten aus vielen verschiedenen Datenquellen erlauben. Mit SemaPlorer, basierend auf KAT und NetworkedGraphs, haben wir dies erreicht und sorgen für eine facettierte, interaktive Suche und Visualisierung über einen sehr großen Satz von semantisch heterogenen und verteilten Daten von unterschiedlicher Qualität. Zwar existieren verschiedene Systeme, die hoch skalierbares Management von RDF-Daten ermöglichen, z.B. YARS2 (Harth et al. 2007). Diese Systeme zielen jedoch auf die Steuerung eines großen Volumens von RDF-Daten in einem einzigen, wenn auch möglicherweise hardwaremäßig verteilten Repository ab und nicht auf die Verknüpfung mehrerer verteilter Repositories, wie die für SemaPlorer verwendete Infrastruktur.

Im Gegensatz dazu zielt unsere Infrastruktur auf die Integration von mehreren semantisch heterogenen Repositories im Sinne des Semantic Web in eine einzige virtuelle Repository-Infrastruktur. DARQ, eine Erweiterung des leichtgewichtigen und in PHP geschriebenen SPARQL-Servers ARQ (http://arc.semsol.org), ist ein verwandter Ansatz zur Abfrage mehrerer SPARQL-Endpunkte (Quilitz u. Leser 2008). Im Gegensatz zu unserem System basiert es auf der Grundlage von manuell gepflegten Statistiken über die verteilten Endpunkte, bei denen wir nicht davon ausgehen, dass sie zur Verfügung stehen. Darüber hinaus werden durch die Struktur der Anfragen von DARQ große Beschränkungen auferlegt. Im Rahmen der Linked Open Data Bemühungen, ergeben sich Herausforderungen ähnlich zu unseren in Bezug auf die Speicheranforderungen. Allerdings konzentriert sich die Linked Open Data Initiative auf das Browsing der Daten und ermöglicht keine komplexen Anfragen. Der relationale Ansatz DynaQuest (Grawunder u. Köster 2003) zielt auf eine verteilte virtuelle relationale Datenbank in Web-Größenordnung. Allerdings kommen relationale Datenbanken nicht gut mit semi-strukturierten, semantisch heterogenen Daten zurecht.

Kartenbasierte Anwendungen wie SemaPlorer sollen interaktiv sein und den Benutzer in der Durchführung einfacher Analyseaufgaben unterstützen (Wisniewski et al. 2009). Existierende Evaluationen haben sich dabei auf unterschiedliche Aspekte konzentriert, wie z. B. die Interkation mit einer

Karte auf dem mobilen Endgerät (Wilson et al. 2006), die Navigation in einer kartenbasierten 3D-Umgebung (Swan et al. 2003) oder der Vergleich zwischen einer 2D- und 3D-Kartennavigation (Porathe u. Prison 2008). Zur facettierten, interaktiven Suche und Visualisierung existieren umfangreiche Designempfehlungen basierend auf langjährige Erfahrungen und Evaluationen (Hearst 2006; Wilson et al. 2009). Die Evaluation einer facettierten, kartenbasierten Anwendung wie SemaPlorer, die sich der Verknüpfung sehr großer, semantischer Datenquellen unterschiedlicher Herkunft und Qualität bedient, ist bisher nicht Gegenstand von Evaluationen gewesen.

8 Zusammenfassung

In diesem Artikel haben wir die SemaPlorer-Anwendung und die zugrunde liegende Dateninfrastruktur präsentiert. Wie gezeigt wurde, ist SemaPlorer ein einfach zu bedienendes Werkzeug, dass dem Endnutzer erlaubt, interaktiv sehr große, verteilte, semantische Datenmengen von unterschiedlicher Qualität interaktiv zu explorieren und zu visualisieren. Die Anwendung setzt einen signifikanten Teil der Daten, die für die Billion Triples Challenge 2008 zur Verfügung standen, ein. Darüber hinaus ist ein großer in RDF umgewandelter Flickr-Datensatz einbezogen worden. Die zugrunde liegende Speicherinfrastruktur ermöglicht einen transparenten Zugriff auf beliebige, verteilte RDF-Repositories, in unserem Fall auf Amazon EC2 betrieben. Mit dieser Speicherinfrastruktur ist die Anwendung in Bezug auf die Zahl der verteilten Komponenten skalierbar. Darüber hinaus können zu einem späteren Zeitpunkt beliebige zusätzliche Daten hinzugefügt werden.

Insgesamt kommen wir mit Amazon EC2 und NetworkedGraphs näher an die Vision des generischen Zugangs zu verteilten semantischen Multimedia-Daten. Insbesondere haben wir gezeigt, dass neben der Skalierung von zentralen Repositories die Verbindung vieler kleiner Repositories in vielerlei Hinsicht ein günstiger, machbarer und erfolgversprechender Ansatz ist, um den Anforderungen des Semantic Web und dessen Skalierbarkeit gerecht zu werden. Auf lange Sicht wird es sinnvoll sein, direkt auf die von den Anbietern der Daten angebotenen SPARQL-Endpunkte zuzugreifen. Die Umstellung auf diese Live-Datenquellen können einfach durch Änderung der Konfigurationen im Federator und ohne Änderung der SemaPlorer-Anwendung oder einer anderen Anwendung, die unsere verteilte Dateninfrastrukturen nutzt, bewerkstelligt werden. Insbesondere

wurde inzwischen ein SPARQL-Endpunkt umgesetzt, der Anfragen in Flickr-API-Aufrufe umsetzt. Die Flickr-API erlaubt den Zugriff auf die Bilder und Metadaten von Flickr mit Hilfe eines normalen Java-Programmes. Der SPARQL-Endpunkt konnte im laufenden Betrieb in das System integriert werden, ohne die eigentliche Anwendung zu ändern.

Danksagung. Diese Forschung wurde co-finanziert von der EU im 6. RP in der NoE K-Space (027026) und Neon-Projekt (027595) und dem RP7 im WeKnowIt Projekt (215453).

Literaturverzeichnis

Arndt R, Troncy R, Staab S, Hardman L und Vacura M (2007) COMM: Designing a Well-Founded Multimedia Ontology for the Web. In: ISWC

Grawunder M und Köster F (2003) The DynaQuest-Framework for Dynamic and Adaptive Source Selection. In: Collaborative Technologies and Systems

Harth A, Umbrich J, Hogan A und Decker S (2007) YARS2: A Federated Repository for Querying Graph Structured Data from the Web. In: ISWC, Springer

Hearst M A (2006) Design recommendations for hierarchical faceted search interfaces. In: SIGIR, Workshop on Faceted Search

Hildebrand M, van Ossenbruggen J und Hardman L (2006) /facet: A Browser for Heterogeneous Semantic Web Repositories. In: ISWC

Munroe K D, Ludscher B und Papakonstantinou Y (2000) Blending Browsing and Querying of XML in a Lazy Mediator System. In: Extending Database Technology

Porathe T und Prison J (2008) Design of human-map system interaction. In: Extended abstracts on Human factors in computing systems

Quilitz B und Leser U (2008) Querying Distributed RDF Data Sources with SPARQL. In: ESWC

Schenk S und Petrak J (2008) Sesame RDF Repository Extensions for Remote Querying. In: ZNALOSTI Conf.

Schenk S und Staab S (2008) NetworkedGraphs: a declarative mechanism for SPARQL rules, SPARQL views and RDF data integration on the web. In: WWW

schraefel m c, Smith D A, Owens A et al. (2005) The evolving mspace platform: leveraging the semantic web on the trail of the Memex. In: Hypertext

Swan J E, Gabbard J L, Hix D et al. (2003) A Comparative Study of User Performance in a Map-Based Virtual Environment. In IEEE Virtual Reality

Wilson M, Russell A, schraefel m c und Smith D A (2006) mSpace mobile: a UI gestalt to support on-the-go info-interaction. In: Extended abstracts on Human factors in computing systems

Wilson M L, schraefel m c und White R W (2009) Evaluating Advanced Search Interfaces using Established Information-Seeking Models. In: Journal of the American Society for Information Science and Technology 60(7)

Wisniewski P K, Pala O, Lipford H R et al. (2009) Grounding geovisualization interface design: a study of interactive map use. In: Extended abstracts on Human factors in computing systems

Yee K P, Swearingen K, Li K und Hearst M (2003) Faceted metadata for image search and browsing. In: Human factors in computing systems, ACM

Zemanek J, Schenk S und Svatek V (2008) Optimizing SPARQL queries over disparate RDF data sources through distributed semi-joins. In: ISWC 2008 Poster and Demo Session Proceedings, CEUR-WS statt nur SemaPlorer