

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

Ansgar Scherp, Simon Schenk, Carsten Saathoff und Steffen Staab

## 1 Inhalt

2	1	Hintergrund .....	315
3	2	Motivation .....	318
4	3	SemaPlorer-Anwendung .....	318
5	4	Datensatz und Vernetzung der Daten .....	319
6	5	SemaPlorer-Architektur .....	322
7	6	Evaluation .....	324
8	6.1	Planung der Evaluation .....	325
9	6.2	Durchführung der Evaluation .....	325
10	6.3	Analyse der Evaluationsergebnisse .....	326
11	7	Verwandte Arbeiten .....	328
12	8	Zusammenfassung .....	329
13		Literatur .....	330

## 14 1 Hintergrund

15 Mit dem Übergang vom Web 1.0 zum Web 2.0 ist das Internet und sein Inhalt noch  
 16 einmal spürbar gewachsen: Internetnutzer sehen sich einer kaum zu überschaubaren  
 17 Flut an Informationen, Bildern und Applikationen gegenüber. Vor diesem Hinter-  
 18 grund erwächst auf Seiten von Nutzern aber auch kommerziellen Anbietern der  
 19 Wunsch, Internetinhalte besser handhabbar zu machen. Ein Beispiel: Informationen  
 20 über Reiseziele werden heutzutage selbstverständlich im Internet gesucht. Dazu  
 21 werden zahlreiche Wikis, Portale und Webseiten aufgesucht, die eine unüberschaubare  
 22 Anzahl von Texten, Bildern und Metainformationen enthalten, die von Inter-  
 23 netnutzern online gestellt werden. Diese für den Benutzer schnell, sinnvoll und op-  
 24 tisch ansprechend nutzbar zu machen, ist eine Herausforderung, der mit der von den

---

A. Scherp (✉)

Institute for Web Science and Technologies, University of Koblenz-Landau,  
 Universitaetsstrasse 1, 56070 Koblenz, Deutschland  
 E-Mail: scherp@uni-koblenz.de

25 Autoren entwickelten Java-basierten, Web 2.0 Anwendung SemaPlover Rechnung  
26 getragen wird. Die SemaPlover-Anwendung verknüpft verschiedene, sehr große  
27 Datenquellen unterschiedlicher Herkunft und Qualität auf intelligente Art und Wei-  
28 se und stellt sie dem Benutzer als so genannten Web 2.0 Mashup dar. Eine zentrale  
29 Herausforderung ist dabei die Skalierbarkeit der sehr großen Datenmengen. Anstatt  
30 sich manuell durch Suchmaschinen und Portale begeben zu müssen, zeigen wir mit  
31 SemaPlover eine Möglichkeit auf, die von Internetnutzern zur Verfügung gestellten  
32 Informationen sinnvoll und ihrem inhaltlichen Zusammenhang gemäß zu sortieren,  
33 anwenderbezogen und in einem angemessenen Zeitrahmen in einem einzigen Sys-  
34 tem zu präsentieren und interaktiv erfahrbar zu machen. Die Datengrundlage von  
35 SemaPlover bilden große semantische Datenbestände wie DBpedia, GeoNames,  
36 WordNet und persönliche FOAF-Dateien. Mit FOAF-Dateien können Benutzer ein  
37 Profil im Internet veröffentlichen, wie es mittels bekannter sozialer Netzwerkplatt-  
38 formen möglich ist. Über semantische Beschreibungen sind die einzelnen Daten-  
39 bestände miteinander verknüpft. Sie sind außerdem verbunden mit einem großen  
40 Flickr-Datensatz, der mittels des Resource Description Frameworks (RDF)<sup>1</sup> seman-  
41 tisch beschrieben wird.

42 Ziel dieses Beitrages ist es, die wissenschaftlichen Hintergründe, die Entwick-  
43 lung und den Nutzen der Java-basierten, Web 2.0 Mashup-Anwendung SemaPlover  
44 für den Endanwender vorzustellen. Der Beitrag ist also sowohl aus wissenschaft-  
45 licher Sicht als auch aus Praxisgründen interessant und relevant. So wird mit der  
46 SemaPlover-Anwendung eine der führenden Technologien im Bereich der verteil-  
47 ten Anfrage von semantischen Datenquellen vorgestellt. Der Artikel demonstriert  
48 den Einsatz dieser Technologie in einem konkreten, praxisrelevanten Szenario und  
49 schafft damit den Transfer von der Forschung in die Praxis.

## 50 2 Motivation

51 Das Internet ist eine wichtige Quelle für Informationen über Städte, Urlaubsorte  
52 oder andere interessante Regionen. Heutige Anwendungen, die Nutzern für diese  
53 Aufgabe zur Verfügung stehen, sind zentralisiert und monolithisch, z. B. Reise-  
54 Websites wie Tripadvisor (<http://www.tripadvisor.com>) und Wikitravel (<http://wikitravel.org>) oder Wissensplattformen wie Freebase (<http://www.freebase.com>).  
55 Mit unserer neuartigen Infrastruktur und Web 2.0 Mashup-Anwendung SemaPlover  
56 greifen wir auf ein Netz verbundener Datenbestände zu. Diese sind nahtlos  
57 in einer einzigen verteilten Infrastruktur integriert, um generischen Zugang zu  
58 den semantischen Multimedia-Daten zu erhalten. Die verschiedenen Datenbe-  
59 stände werden über SPARQL<sup>2</sup>-Endpunkte zur Verfügung gestellt. Über solche  
60 Endpunkte können semantische Datenbanken über die Anfragesprache SPARQL  
61

---

<sup>1</sup> Resource Description Framework, <http://www.w3.org/RDF/>

<sup>2</sup> SPARQL Query Language, <http://www.w3.org/TR/rdf-sparql-query/>

62 angesprochen werden. Damit können nahezu beliebige Datenquellen ad hoc zur  
63 Dateninfrastruktur von SemaPlorer hinzugefügt werden. Um Informationen aus  
64 dieser verteilten Infrastruktur abzurufen und zu visualisieren, bedienen wir uns  
65 mit der SemaPlorer-Anwendung dem sogenannten „Blended Browsing and Que-  
66 rying“-Ansatz (Munroe, Ludscher u. Papakanstantinou 2000). Die Nutzer können  
67 sich durch nahezu beliebige Datensätze unter Verwendung verschiedener Ansich-  
68 ten (Facetten) wie Ort, Zeit, Personen und Tags navigieren (Hearst 2006). Wenn  
69 der Benutzer mit der Anwendung interagiert, werden dabei gleichzeitig mehrere  
70 Anfragen an die zugrunde liegende Speicher-Infrastruktur gesendet, um die ent-  
71 sprechenden Ergebnisse zu berechnen. Die Ergebnisse werden mittels einer Kar-  
72 te, Medien und verschiedenen Kontextansichten, die die verschiedenen Facetten  
73 repräsentieren, dargestellt.

74 Für SemaPlorer haben wir verschiedene semantische Datenquellen wie DBpedia  
75 (<http://dbpedia.org>), eine semantische Version von Wikipedia, GeoNames (<http://geonames.org>), eine umfangreiche Datenbank mit geo-referenzierten Orten, Word-  
76 Net (<http://wordnet.princeton.edu>) mit einer Abbildung des englischen Sprachvoka-  
77 bulars und persönliche FOAF-Dateien aus der semantischen Suchmaschine Swoogle  
78 (<http://swoogle.umbc.edu>) integriert. Darüber hinaus haben wir einen partiellen  
79 Crawl, also eine partielle lokale Kopie von Flickr (<http://flickr.com>) erstellt und als  
80 einen sehr großen, nicht-semantischen Datensatz, der umgerechnet auf 700 Mio.  
81 RDF Triple kommt, eingebunden. Der Datensatz umfasst alle Annotationen von  
82 Fotos auf Flickr von ca. Mai 2005 bis April 2006. Zusammen bilden diese Datenbe-  
83 stände einen sehr großen, semantisch heterogenen Datensatz von gemischter Qua-  
84 lität, die zusammen über eine Milliarde Triples ergeben. Die Verknüpfung dieser  
85 Daten erfordert eine flexible und skalierbare Speicherstruktur. Die SemaPlorer-In-  
86 frastruktur besteht aus 25 RDF-Datenbanken. Die Datenbanken werden in virtuel-  
87 len Maschinen auf Amazons Elastic Computing Cloud (EC2, <http://aws.amazon.com/ec2>) gehostet. Die EC2 ist ein Dienst von Amazon, um eigene Anwendungen  
88 im Internet auszuführen und anzubieten. Der Simple Storage Service von Amazon  
89 (S3, <http://aws.amazon.com/s3>) wird genutzt, um die semantischen Datensätze zu  
90 speichern. Er stellt eine zu EC2 passende Infrastruktur zur Verfügung um große  
91 Datenmengen über das Internet bereitzustellen. Die Speicher können wie ein ein-  
92 ziger, virtueller RDF Speicher über einen Federator angesprochen werden. Der Fe-  
93 derator verwendet die von den Autoren entwickelte Technologie NetworkedGraphs  
94 (Schenk u. Staab 2008), einen SPARQL-basierten, verteilten View-Mechanismus  
95 für RDF und verteilte Auswertung von SPARQL-Anfragen (Schenk u. Petrak 2008;  
96 Zemaniek, Schenk u. Svatek 2008). NetworkedGraphs erlaubt einfaches, regelba-  
97 siertes Schließen zur Laufzeit, zum Beispiel für die Integration semantisch hetero-  
98 gener Daten. Das Verteilen von Anfragen innerhalb der Infrastruktur wird durch  
99 eine – ebenfalls RDF-basierte – Konfiguration gesteuert, die im Repository des  
100 Federators gespeichert ist. Diese Konfiguration kann zur Laufzeit des Systems an-  
101 gepasst werden. Daher wird das Hinzufügen neuer Datenquellen durch die Anpas-  
102 sung der Federator-Einstellungen extrem einfach, während sie für die SemaPlorer-  
103 Anwendung vollkommen transparent ist.  
104  
105

### 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 beispielsweise 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 solcher 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 Abb. 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 Abb. 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 Abb. 1 mit Kontextinformationen füllen.

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 Kontext-



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

149 ansichten navigieren können. Zum Beispiel können die Benutzer, wenn die Karte in  
 150 SemaPlorer die Stadt Berlin zeigt, auf den Tag „street art“ (Straßenkunst) klicken.  
 151 Sofort wird die Kartenansicht aktualisiert und die Standorte der Flickr Fotos welche  
 152 mit „streetart“ annotiert sind, angezeigt. Durch die Eingabe einer Suche nach Paris  
 153 kann der Nutzer zwischen dem aktuellen Kontext, nämlich Straßenkunst in Berlin,  
 154 zu Straßenkunst in Paris wechseln und miteinander vergleichen.

## 155 4 Datensatz und Vernetzung der Daten

156 Um die facettrierte, interaktive Suche und Visualisierung in SemaPlorer zu unter-  
 157 stützen, werden verschiedene Arten von semantischen Daten kombiniert. Wir ver-  
 158 wenden einen signifikanten Teil der Datensätze, die für die Billion Triples Chal-  
 159 lenge<sup>3</sup> zur Verfügung gestellt worden sind, nämlich DBpedia (120 Mio. Triples),

<sup>3</sup> 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.

160 GeoNames (70 Mio. Triples), WordNet (2 Mio. Triples) und Swoogle (175 Mio.  
161 Triples). Darüber hinaus verwenden wir einen großen Datensatz von Flickr, der  
162 über mehrere Monate in den Jahren 2005–2006 gesammelt und in RDF übersetzt  
163 wurde (700 Mio. Triples). Wie in Abschn. 2 beschrieben, haben wir verschiedene  
164 Kontextansichten für SemaPlorer definiert. Diese Kontextansichten ergeben sich  
165 aus den Eigenschaften beziehungsweise den zur Verfügung gestellten Informatio-  
166 nen der verwendeten Daten. Im Folgenden beschreiben wir die verwendeten Daten  
167 entlang der vier in SemaPlorer definierten Facetten und wie sie miteinander ver-  
168 bunden sind.

169 **Ort** Elemente dieser Facette beziehen sich auf die geographischen Koordinaten.  
170 Wir setzen GeoNames für Orte aller Art ein wie beispielsweise Städte, Länder und  
171 andere. Für Sehenswürdigkeiten verwenden wir eine Kombination der Volltext-  
172 Suche auf Artikelbeschreibungen aus DBpedia und deren Kategoriebeschreibun-  
173 gen, welche mit dem SKOS-Vokabular (Simple Knowledge Organization System,  
174 <http://www.w3.org/2004/02/skos>) beschrieben sind. Mit SKOS können Systeme  
175 zur Wissensorganisation beschrieben werden wie Thesauri, Klassifikationssche-  
176 mas und Taxonomien. Zur Erkennung von Sehenswürdigkeiten betrachten wir die  
177 SKOS-Kategorien in DBpedia, insbesondere das Konzept SKOS:broader welches  
178 hierarchische Beziehungen zwischen SKOS Konzepten beschreibt und berechnen  
179 die transitive Hülle aller Kategorien. Außerdem nutzen wir eine Volltext-Suche  
180 auf den Kategorienamen und schränken die Ergebnisse auf Einträge ein, die in der  
181 Kategorie `dbpedia:Visitor_attractions` einsortiert sind. Für die Anzeige der nahege-  
182 legenen Orte und Sehenswürdigkeiten wählen wir alle Geschwister eines gewählten  
183 Standortelements und sortieren sie auf der Grundlage der geografischen Distanz.  
184 Wenn z. B. der Arc de Triomphe in Paris ausgewählt wurde, werden als nahe Orte  
185 der Eiffelturm und Notre Dame angezeigt. Zusätzlich werden Bilder von Flickr  
186 dargestellt, die mit Geoinformationen versehen sind und sich im relevanten Karten-  
187 ausschnitt befinden.

188 **Zeit** Für die Zeit-Facette sind keine expliziten Daten definiert. Stattdessen können  
189 Inhalte aus einem bestimmten Zeitraum ausgewählt werden, z. B. Bilder aus einem  
190 bestimmten Monat aus Flickr. Darüber hinaus können Inhalte nach bestimmten Jah-  
191 reszeiten wie Sommer und Winter gefiltert werden. Die Zeit-Facette ist nicht für  
192 die SemaPlorer-Anwendung implementiert worden und als zukünftige Erweiterung  
193 vorgesehen.

194 **Person** In den von SemaPlorer verwendeten Datensätzen haben wir drei Arten von  
195 Personen identifiziert: Diese sind Persönlichkeiten aus DBpedia, Flickr-Benutzer  
196 die Bilder eingestellt haben und Internet-Nutzer, die ihre FOAF-Dateien veröffent-  
197 licht haben und über Swoogle zugreifbar sind. Für jede dieser Kategorien von Per-  
198 sonen verwenden wir eine andere Kombination der Daten. Für Persönlichkeiten  
199 wählen wir Bilder, die die ausgewählten Persönlichkeiten zeigen, basierend auf  
200 einer Volltext-Suche auf den Flickr-Tags. In Bezug auf einen Flickr-Nutzer suchen  
201 wir nach Inhalten, die durch den Benutzer veröffentlicht wurden. Für Internet-Nut-  
202 zer betrachten wir den Geostandort in der FOAF-Datei (falls vorhanden) und ver-  
203 binden sie mit Bildern von diesem Ort aus Flickr.

204 **Tags** Tags stehen direkt im Zusammenhang mit den Flickr-Inhalten. Wir bieten  
205 Volltextsuche über die Tags. Wenn ein Tag von einem Nutzer ausgewählt wurde,  
206 zeigen wir verwandte Tags von Flickr sowie WordNet. Hinsichtlich Flickr sind dies  
207 alle Tags der aktuell dargestellten Photos. Verwandte Tags in WordNet sind die Syn-  
208 onyme des aktuellen Tag.

209 **Komplexität der Anfragen** Um die oben beschriebenen Facetten mit Inhalten zu  
210 füllen werden mehrere Anfragen gleichzeitig ausgeführt. Für die initiale Suche mit-  
211 tels Stichworten, wie in Abschn. 3 beschrieben, werden gleichzeitig drei Abfragen  
212 nach Orten, Personen und Tags durchgeführt. Bei einem Klick auf einen der gefun-  
213 denen Einträge in der Ergebnisliste werden acht gleichzeitige Anfragen ausgeführt,  
214 um die Medien- und die Kartenansicht zu füllen, die nahegelegenen Orte zu berech-  
215 nen, Sehenswürdigkeiten, Prominente und Flickr-Benutzer, Internet-Nutzer und  
216 Tags auszuwählen und die Zusammenfassung aus DBpedia zu erhalten. Die glei-  
217 chen Anfragen werden durchgeführt, wenn der Kontext mit der aktuellen Ansicht  
218 geändert wird, z. B. wenn der Standort geändert wird, indem auf ein Bild oder einen  
219 Ort in der Nähe oder eine spezielle Person geklickt oder ein Tag in der entspre-  
220 chenden Kontextansicht gewählt wird. Dieser Ansatz ermöglicht eine sehr flexible  
221 Änderung der SemaPlorer-Anwendung um z. B. bestimmte Elemente der Ansichten  
222 hinzuzufügen oder zu entfernen. Die Anfragen nutzen die volle Ausdruckskraft von  
223 SPARQL. Darüber hinaus ermöglichen wir über die Textsuchmaschine Lucene bzw.  
224 LuceneSail eine Volltextsuche in SPARQL (<http://dev.nepomuk.semanticdesktop.org/wiki/LuceneSail>). Wir haben das LuceneSail erweitert, um Anfragen zu unter-  
225 stützen, die die Form „A ODER B“ haben, und Anfragen nach geographischer Nähe  
226 zu ermöglichen. Die Anfragen bestehen typischerweise aus vier bis neun Joins, das  
227 heißt im Durchschnitt verbinden sie also bis zu vier Datensätze in einer einzigen  
228 Abfrage. Da die GeoNames- und Flickr-Datensätze über mehrere Repositories ver-  
229 teilt sind, werden für diese Datensätze jeweils mehrere Anfragen ausgeführt. Diese  
230 sind jedoch unkritisch, da sie leicht parallelisiert werden können. Je nach Kontext,  
231 den der Benutzer wählt, können die Anfragen komplexer werden, z. B. indem der  
232 Benutzer die Bilder auswählt, die mehrfach getagt sind und zusätzlich räumlich  
233 eingeschränkt sind.  
234

235 **Erfolge und Erfahrungen** Bei der Erstellung des Datensatzes für unsere Sema-  
236 plorer-Anwendung haben wir bemerkt, dass die Datensätze oft nicht vollständig  
237 und manchmal auch bezüglich der Semantik nicht eindeutig genug sind: Zum Bei-  
238 spiel fehlen in GeoNames zu einem beliebigen Eintrag Informationen über Sehens-  
239 würdigkeiten und Orte in der Nähe – Informationen, die in der HTML-Version  
240 vorhanden sind. Trotzdem konnten wir diese Informationen durch die Verbindung  
241 der einzelnen Datensätze, wie oben beschrieben, gewinnen. Des Weiteren haben wir  
242 beobachtet, dass die Daten auch innerhalb eines einzelnen Datensatzes heterogen  
243 sind. Zum Beispiel gibt es keinen klaren Lösungsansatz für die Angabe des Geburts-  
244 ortes einer Person in DBpedia. Manchmal ist es `dbpedia:cityofbirth` und manch-  
245 mal `dbpedia:placeofbirth`. In SemaPlorer lösen wir diese Unklarheiten durch die  
246 Zusammenfassung der beiden Eigenschaften in einem View. Ein View erlaubt eine  
247 bestimmte Sicht auf einen Datensatz zu legen und ermöglicht somit die Vereinheit-

248 lichung der beiden Modellierungsvarianten der DBpedia durchzuführen. Während  
249 Linked Open Data, also die Verknüpfung von semantischen Datenbeständen, fort-  
250 schreitet, ist es immer noch eine offene Frage, wie es für die Verwaltung von Res-  
251 sourcen wie Flickr-Bilder zu nutzen ist. Wie SemaPlorer zeigt, ist eine Kartierung  
252 der Linked Open Data und die semantische Beschreibung der Flickr-Daten in RDF  
253 möglich und funktioniert z. B. mit GeoNames gut. Doch statt der Kennzeichnung  
254 von Bildern mit Tags und anschließender Kartierung dieser Tags in Zusammenhang  
255 mit Open Data wäre es gewinnbringender, direkt Linked Open Data zur Annotation,  
256 also Beschreibung der Bilder mittels semantischer Konzepte zu verwenden. Zum  
257 Beispiel könnte ein Bild, das den Eiffelturm zeigt, direkt mit dem entsprechenden  
258 Konzept für den Eiffelturm aus der DBpedia annotiert werden.

## 259 5 SemaPlorer-Architektur

260 Die Architektur der SemaPlorer-Anwendung und Infrastruktur ist in Abb. 2 darge-  
261 stellt. Sie gliedert sich in zwei Sub-Systeme: Das erste Sub-System besteht aus dem  
262 K-Space Annotation Tool (KAT, <https://launchpad.net/kat>) und seinen SemaPlorer-  
263 spezifischen Erweiterungen, den KAT-Plugins. Es wird auf den Client-Computern  
264 eingesetzt und bietet die Benutzerschnittstelle und die Anwendungslogik von Se-  
265 maPlorer. Das zweite Sub-System implementiert die verteilte Dateninfrastruktur  
266 und eine administrative Komponente für RDF-Repositories. Dies beinhaltet den  
267 NetworkGraphs-basierenden Federator und die verschiedenen RDF-Repositories  
268 für die semantischen Daten für die DBpedia-Abstracts und Flickr-Tags. Die Ver-  
269 waltung der Komponenten und der Federator werden auf unserer EDV-Infrastruktur  
270 gehostet. Alle anderen Komponenten werden auf Amazon EC2 Knoten gehostet.  
271 Die Architektur von SemaPlorer und die einzelnen Komponenten werden im Detail  
272 im Folgenden beschrieben.

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

283 Der in Abschn. 3 beschriebene Datensatz wird durch das zweite Subsystem, die  
284 auf NetworkedGraphs basierende verteilte Dateninfrastruktur unter Einsatz von  
285 Amazon EC2 realisiert. Die Verwaltungskomponente (Administration Component)  
286 dieser Dateninfrastruktur kontrolliert die virtuellen Maschinen, die auf EC2 laufen.  
287 Über eine einfach zu bedienende Web-GUI, können EC2-Knoten für spezielle Tei-  
288 le der Daten oder der gesamte Datenbestand gestartet und gestoppt werden. Neue

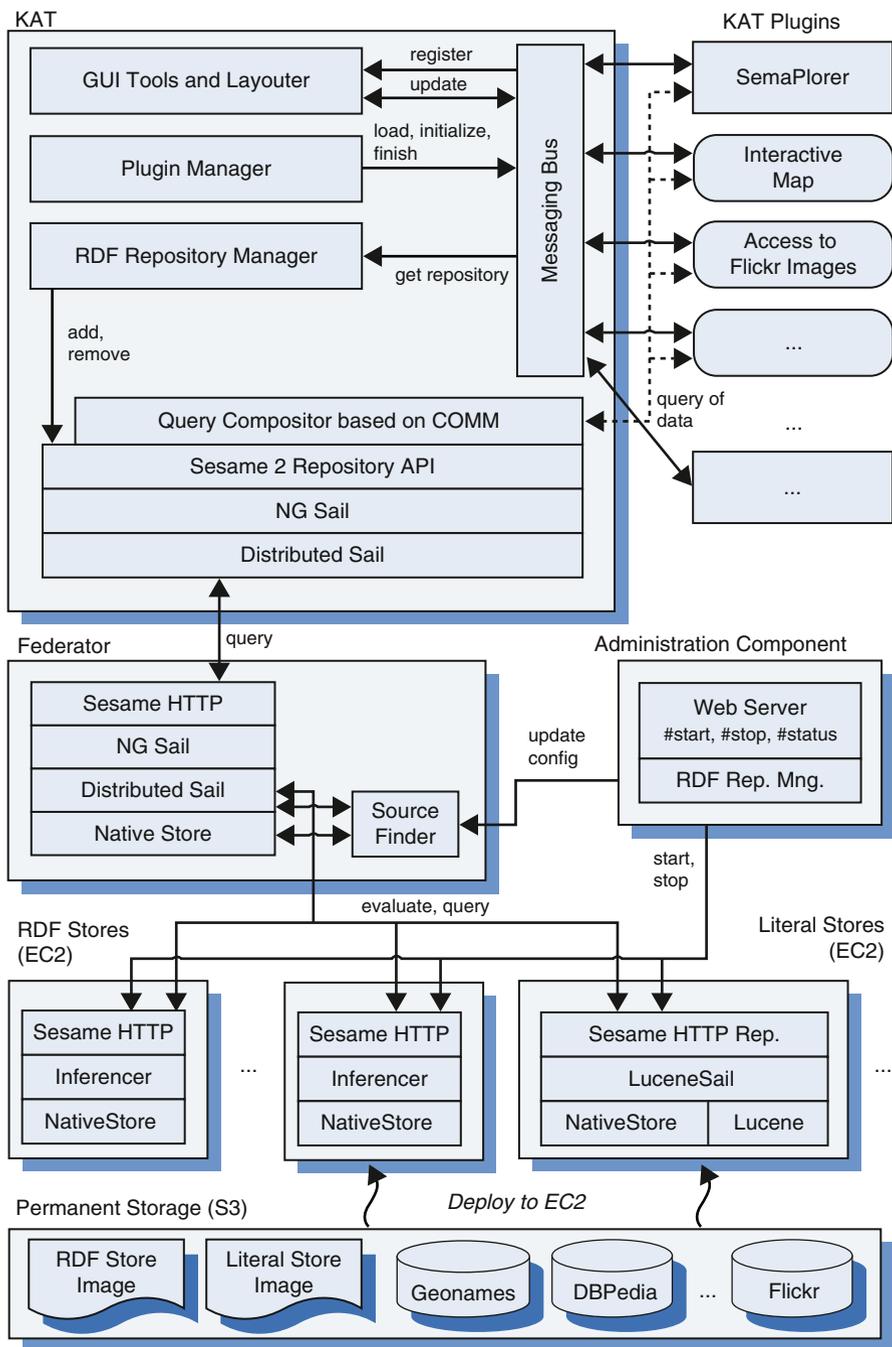


Abb. 2 Architektur von SemaPlover

289 Datensätze können durch Hinzufügen einer Beschreibung des Datensatzes zu einer  
290 Konfigurationsdatei und den Start des neuen Knotens geschaffen werden. Wenn  
291 Knoten gestartet oder gestoppt werden, aktualisiert die Verwaltungskomponente  
292 die Federator-Konfiguration. Der Federator ist der einzige SPARQL-Endpunkt, den  
293 SemaPlorer direkt nutzt, und verbirgt die Komplexität der unterliegenden Infra-  
294 struktur. Anfragen an den Federator werden analysiert, um festzustellen, welche  
295 Endpunkte zur Auswertung der Anfragen genutzt werden können. Anschließend  
296 wird die Anfrage in Unterabfragen aufgeteilt, die in den jeweiligen Repositories  
297 ausgewertet werden (Schenk u. Petrak 2008; Zemanek, Schenk u. Svatek 2008).

298 Die Datensätze werden dabei in Speicherknoten der EC2 mittels S3 gespeichert.  
299 Wir nutzen drei verschiedene Konfigurationen für EC2 Knoten: Die erste speichert  
300 RDF-Daten ohne jegliche Folgerungen. Sie wird z. B. für DBpedia Infobox Da-  
301 ten verwendet. Sie dient auch als Grundlage für die anderen beiden Knotentypen.  
302 Die zweite verwendet das LuceneSail und bietet zusätzlich Volltextindizes über  
303 die RDF-Literale. Es wird z. B. für die Tags, DBpedia Artikel- und Kategorienna-  
304 men verwendet. Für die SemaPlorer-Anwendung brauchen wir kein vollständiges  
305 RDFS-Reasoning. Im Gegensatz dazu wird die Transitivität in SKOS-Hierarchien  
306 benötigt, die nicht in RDFS geboten wird. Daher verwenden wir benutzerdefinierte  
307 Regeln in der dritten Konfiguration von S3-Knoten. Da der entsprechende Infe-  
308 renzer in Sesame nicht für den verwendeten Datensatz skaliert, führen wir eine  
309 Vorberechnung der transitiven Hülle von SKOS:broader für DBpedia-Kategorien  
310 durch. Vereinfacht gesagt berechnen wir also eine Erweiterung der SKOS:broader  
311 Relation, die zusätzlich alle indirekt erreichbaren Paare enthält (und damit transitiv  
312 ist). Der Federator erlaubt zudem die Definition von einfachen Views, um homo-  
313 gene Darstellungen aus den verschiedenen Datenquellen anzubieten. Ein Beispiel  
314 ist das oben genannte Vokabular für Geburtsorte in DBpedia. Dieses Schema-Ma-  
315 pping erfolgt während der Laufzeit mit NetworkedGraphs. Zum Beispiel haben wir  
316 für „Personen“ drei verschiedene Darstellungen: FOAF-Dateien, die das FOAF  
317 Vokabular-benutzen, DBpedia mit einer Personenkategorie und Flickr-Benutzer.  
318 Ähnliche Herausforderungen ergeben sich aus der Modellierung von räumlichen  
319 Einheiten und Annotationen von Bildern und bei Einträgen ohne ein klares Sche-  
320 ma, wie der Geburtsort in DBpedia. Um SemaPlorer zu ermöglichen, von diesen  
321 verschiedenen Darstellungen zu abstrahieren, gestalten wir sie in einer kanonischen  
322 Form. Im Falle von Personen wird das FOAF-Vokabular verwendet. Als Ergebnis  
323 können wir jeden Datensatz-hinzufügen, der das FOAF-Vokabular einsetzt.

## 324 6 Evaluation

325 Ziel der Billion Triples Challenge ist es, die Skalierbarkeit von Semantischen Web-  
326 technologien auf mehr als eine Milliarden Tripeln zu demonstrieren und damit et-  
327 was Sinnvolles zu tun. Als solches wurde die Java-basierte, Web 2.0 Mashup-An-  
328 wendung SemaPlorer und ihre zu Grunde liegende Infrastruktur als ein technischer  
329 Demonstrator, aber nicht als eine Endbenutzer-Anwendung, die in einer echten pro-

330 duktiven Umgebung läuft, entworfen. Um in einer solch frühen Phase eine Rück-  
331 meldung über die Benutzbarkeit und Nützlichkeit der Anwendung und Verbesse-  
332 rungsvorschläge zu erhalten, wurde eine formative Evaluation durchgeführt. Wir  
333 baten 20 Personen aus dem Institut für Informatik der Universität Koblenz-Landau  
334 (11 Doktoranden, 9 Studierende), SemaPlorer auszuprobieren. Die Personen sind  
335 zwischen 21 und 26 Jahren alt und haben gute bis sehr gute Computer-Kenntnis-  
336 se. 18 Teilnehmer haben bereits Erfahrung mit der Nutzung von Karten-basierten  
337 Anwendungen zur Informationsbeschaffung und Visualisierung wie z. B. Google-  
338 Maps. Sie verwenden diese Anwendungen für Reiseplanungen (75 %), um Informa-  
339 tionen über den Ort zu erhalten (55 %) und für berufliche Zwecke (25 %). Demzu-  
340 folge sind die Testkandidaten typische Benutzer von Anwendungen wie SemaPlorer  
341 und sind gute Kandidaten zur Ermittlung relevanter Rückmeldungen.

## 342 **6.1 Planung der Evaluation**

343 Die Evaluation der SemaPlorer-Anwendung wurde in drei Phasen unterteilt, näm-  
344 lich Einführung, Test und Rückmeldung. In der Einführungsphase wurden die Teil-  
345 nehmer mit SemaPlorer und seinen Features vertraut gemacht. Den Teilnehmern  
346 wurde erklärt, dass es nicht um die Messung ihrer Leistungen bei der Abarbeitung  
347 der Evaluationsaufgaben geht, sondern um die Gewinnung von Erkenntnissen zur  
348 Verbesserung von SemaPlorer. In der nachfolgenden Testphase erfolgte die eigent-  
349 liche Bewertung. Jeder Teilnehmer hatte eine festgelegte Zahl von Aufgaben aus-  
350 zuführen. Eine solch einheitliche Aufgabenstellung ist wichtig, um eine Vergleich-  
351 barkeit zwischen den einzelnen Testpersonen herstellen zu können und eine valide  
352 Rückmeldung zu erhalten. In der Feedback-Phase füllten die Teilnehmer einen Fra-  
353 gebogen aus. Die Fragen zur Erfassung der Zufriedenheit der Teilnehmer haben  
354 wir in Anlehnung an den IsoMetrics-L Fragebogen erstellt. Es wurde jedoch keine  
355 explizite Gewichtung der einzelnen Fragen vorgenommen, sondern den Benutzern  
356 die Möglichkeit gegeben, punktuell subjektive Rückmeldungen zu den einzelnen  
357 Fragen bzw. Funktionalitäten der Anwendung die sie für wichtig erachten zu geben.  
358 Für die Testphase hatten die Teilnehmer keine Zeitbegrenzung, sie konnten sich so  
359 viel Zeit lassen, wie sie benötigen, um die Aufgaben zu erfüllen.

## 360 **6.2 Durchführung der Evaluation**

361 Die Dauer der Sitzungen lag zwischen 10 und 60 min (Durchschnitt 30, Median  
362 25). Demzufolge haben die Teilnehmer eine angemessene Zeit mit der Lösung der  
363 Aufgaben verbracht. Die Aufgaben, die ausgeführt werden sollten, sind die Suche  
364 nach der Stadt Berlin und die Verwendung des „Sights“-Features gewesen. Dann  
365 sollte das Ergebnis durch Hinzufügen des „streetart“-Tags auf die Anzeige von Bil-  
366 dern 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.

### 6.3 Analyse der Evaluationsergebnisse

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 einzelnen Features der SemaPlorer-Anwendung sind im Fragebogen im Durchschnitt mit Werten zwischen 0,9 bis 3,3 beurteilt worden. Wir erklären uns diese

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

Frage	Mittelwert	Standardabweichung
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 Antwortzeiten der Anwendung entsprechen meinen Erwartungen	2,5	1,2

Bewertung in den unteren beiden Dritteln der Skala durch die Heterogenität der Daten, die für die SemaPlorer-Anwendung genutzt wurden, die Performanz der Anwendung und die Benutzbarkeit. Die für die SemaPlorer-Anwendung verwendeten Daten stammen aus unterschiedlichen Quellen und sind von unterschiedlicher Qualität. GeoNames, DBpedia und der Flickr-Datensatz sind durch die Mitwirkung einer großen Anzahl an Benutzern entstanden. Für solche Datensätze kann die Qualität der Anfrageergebnisse nicht garantiert werden und ist stark von der jeweiligen individuellen Anfrage abhängig. Dies spiegelt sich in unserer Evaluation wider, dass die Nutzer die Qualität der Suchergebnisse als mittelmäßig einstufen (S1). Die Aufteilung der Suchergebnisse in Orte, Tags und Personen wurde ähnlich bewertet (S2). Der Kontextwechsel durch die Nutzung der Suchfunktion sollte intuitiver gestaltet werden (S3). Die Benutzbarkeit von Karten- und Medienansicht wurde als durchschnittlich beurteilt (A1 und A2). Bei der Beurteilung der einzelnen Features der Facetten wurde die Auswahl der Sehenswürdigkeiten in der Ort-Facette am besten bewertet (F1). Es wurden auch interessante Ansichten gefunden (F2). Das „nearby places“-Feature wurde ähnlich gut wie die Auswahl der Sehenswürdigkeiten bewertet (F3). Allerdings sollte die Qualität der gefundenen „nearby places“ verbessert werden (F4). Dazu benötigen wir bessere Daten über benachbarte Orte als uns bisher zur Verfügung stehen. Die Navigation entlang von WordNet (F5) 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).

In der letzten Phase unserer Evaluation konnten die Testpersonen zusätzliches 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 dienen. 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

472 Open Data Initiative auf das Browsing der Daten und ermöglicht keine komplexen  
473 Anfragen. Der relationale Ansatz DynaQuest (Grawunder u. Köster 2003) zielt auf  
474 eine verteilte virtuelle relationale Datenbank in Web-Größenordnung. Allerdings  
475 kommen relationale Datenbanken nicht gut mit semi-strukturierten, semantisch he-  
476 terogenen Daten zurecht.

477 Kartenbasierte Anwendungen wie SemaPlorer sollen interaktiv sein und den Be-  
478 nutzer in der Durchführung einfacher Analyseaufgaben unterstützen (Wisniewski  
479 et al. 2009). Existierende Evaluationen haben sich dabei auf unterschiedliche As-  
480 pekte konzentriert, wie z. B. die Interaktion mit einer Karte auf dem mobilen End-  
481 gerät (Wilson et al. 2006), die Navigation in einer kartenbasierten 3D-Umgebung  
482 (Swan et al. 2003) oder der Vergleich zwischen einer 2D- und 3D-Kartennavigation  
483 (Porathe u. Prison 2008). Zur facettierten, interaktiven Suche und Visualisierung  
484 existieren umfangreiche Designempfehlungen basierend auf langjährige Erfahrun-  
485 gen und Evaluationen (Hearst 2006; Wilson et al. 2009). Die Evaluation einer fa-  
486 cettierten, kartenbasierten Anwendung wie SemaPlorer, die sich der Verknüpfung  
487 sehr großer, semantischer Datenquellen unterschiedlicher Herkunft und Qualität be-  
488 dient, ist bisher nicht untersucht worden.

## 489 8 Zusammenfassung

490 In diesem Artikel haben wir die SemaPlorer-Anwendung und die zugrunde liegende  
491 Dateninfrastruktur präsentiert. Wie gezeigt wurde, ist SemaPlorer ein einfach zu be-  
492 dienendes Werkzeug, das dem Endnutzer erlaubt, interaktiv sehr große, verteilte,  
493 semantische Datenmengen von unterschiedlicher Qualität interaktiv zu explorieren  
494 und zu visualisieren. Die Anwendung setzt einen signifikanten Teil der Daten, die  
495 für die Billion Triples Challenge 2008 zur Verfügung standen, ein. Darüber hinaus  
496 ist ein großer in RDF umgewandelter Flickr-Datensatz einbezogen worden. Die zu-  
497 grunde liegende Speicherinfrastruktur ermöglicht einen transparenten Zugriff auf  
498 beliebige, verteilte RDF-Repositories, in unserem Fall auf Amazon EC2 betrieben.  
499 Mit dieser Speicherinfrastruktur ist die Anwendung in Bezug auf die Zahl der ver-  
500 teilten Komponenten skalierbar. Darüber hinaus können zu einem späteren Zeit-  
501 punkt beliebige zusätzliche Daten hinzugefügt werden.

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

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

517 **Danksagung** Diese Forschung wurde co-finanziert von der EU im 6. RP in der NoE K-Space  
518 (027026) und Neon-Projekt (027595) und dem RP7 im WeKnowIt Projekt (215453).

## 519 Literatur

- 520 Arndt R, Troncy R, Staab S, Hardman L, Vacura M (2007) COMM: designing a well-founded  
521 multimedia ontology for the web. ISWC, Springer, Berlin, S 30–43
- 522 Grawunder M, Köster F (2003) The dynaquest-framework for dynamic and adaptive source selec-  
523 tion. Collaborative technologies and systems
- 524 Harth A, Umbrich J, Hogan A, Decker S (2007) YARS2: A federated repository for querying graph  
525 structured data from the web. ISWC, Springer, Berlin, S 211–224
- 526 Hearst MA (2006) Design recommendations for hierarchical faceted search interfaces. SIGIR,  
527 workshop on faceted search
- 528 Hildebrand M, van Ossenbruggen J, Hardman L (2006) /facet: A browser for heterogeneous se-  
529 mantic web repositories. ISWC, Springer, Berlin, S 272–285
- 530 Munroe KD, Ludscher B, Papakonstantinou Y (2000) Blending browsing and querying of XML in  
531 a lazy mediator system. Extending database technology
- 532 Porathe T, Prison J (2008) Design of human-map system interaction. Extended abstracts on human  
533 factors in computing systems, ACM, New York, S 2859–2864
- 534 Quilitz B, Leser U (2008) Querying distributed RDF data sources with SPARQL. ESWC, Sprin-  
535 ger, Berlin
- 536 Schenk S, Petrak J (2008) Sesame RDF repository extensions for remote querying. ZNALOSTI  
537 Conf.
- 538 Schenk S, Staab S (2008) Networkedgraphs: a declarative mechanism for SPARQL rules, SPARQL  
539 views and RDF data integration on the web. WWW, ACM, New York, S 585–594
- 540 Schraefel MC, Smith DA, Owens A et al. (2005) The evolving myspace platform: leveraging the  
541 semantic web on the trail of the Memex. Hypertext, ACM, New York, S 174–183
- 542 Swan JE, Gabbard JL, Hix D et al. (2003) A comparative study of user performance in a map-  
543 based virtual environment. In IEEE virtual reality, IEEE Computer Society, Washington, S 259
- 544 Wilson M, Russell A, Schraefel MC, Smith DA (2006) mSpace mobile: a UI gestalt to support  
545 on-the-go info-interaction. Extended abstracts on human factors in computing systems, ACM,  
546 New York, S 247–250
- 547 Wilson ML, Schraefel MC, White RW (2009) Evaluating advanced search interfaces using esta-  
548 blished information-seeking models. J Am Soc Inf Sci Technol 60(7):1407–1422
- 549 Wisniewski PK, Pala O, Lipford HR et al. (2009) Grounding geovisualization interface design:  
550 a study of interactive map use. Extended abstracts on human factors in computing systems,  
551 ACM, New York, S 3752–3762
- 552 Yee KP, Swearingen K, Li K, Hearst M (2003) Faceted metadata for image search and browsing.  
553 Human factors in computing systems, ACM, New York, S 401–408
- 554 Zemanek J, Schenk S, Svatek V (2008) Optimizing SPARQL queries over disparate RDF data  
555 sources through distributed semi-joins. ISWC 2008 poster and demo session proceedings,  
556 CEUR-WS