Cadenza Zugang

Neue Ansätze zur benutzerfreundlichen Suche nach strukturierten Umweltdaten –

Ein begriffsbasierter Einstieg in Cadenza

A. Abecker; W. Kazakos; C. Hofmann; A. Valikov; G. Nagypal
disy Informationssysteme GmbH
Erbprinzenstr. 4-12
76133 Karlsruhe

R. Nedkov; V. Bicer
FZI Forschungszentrum Informatik
Haid-und-Neu-Str. 10-14
76131 Karlsruhe

1. Einleitung

Mit der zunehmenden Umsetzung der INSPIRE-Direktive zum Aufbau einer europäischen Geodateninfrastruktur werden der Öffentlichkeit im großen Maßstab aufwändig erhobene Umweltdaten verfügbar gemacht, deren wertschöpfende Nutzung durch Bürger und Firmen aber offensichtlich noch deutlich ausbaufähig ist. Ein möglicher Grund hierfür ist die schiere Masse und Komplexität der Datenbestände, was dem Laien (und sogar Mitarbeitern der öffentlichen Verwaltung, die außerhalb ihres eigenen Fachgebiets recherchieren) den Zugang sehr schwer macht. Für die Suche in bzw. Nutzung von Daten braucht man Wissen über (1) die Existenz der Daten, (2) ihre relationale Struktur und die technischen Möglichkeiten zu Anfrage, Verarbeitung und Darstellung sowie (3) die Fachbegriffe, die als Tabellen- und Attributnamen, Attributwerte usw. verwendet werden. disy Cadenza vereinfacht die technischen Aspekte (2); dennoch ist Wissen zur Fachlichkeit (1, 3) unabdingbar für die effektive Suche und Nutzung.

Semantische Technologien versprechen Unterstützung bei Suche, Integration und Verarbeitung von Daten, Dokumenten und Diensten, die über Web-Protokolle erreichbar sind /1/. Indem man Web-basierte Informationsquellen mit maschinenlesbaren *Metadaten* versieht, macht man sie besser auffindbar und interpretierbar. Semantische Metadaten nutzen sogenannte Ontologien, reichhaltige Modelle der Begriffsstrukturen und -zusammenhänge in einem Anwendungsgebiet /1/. Dieses Hintergrundwissen eines Fachgebiets kann zur Wissensorganisation genutzt werden und so die Suche und Navigation nach Informationen erleichtern. Ontologien für die Wissensorganisation beinhalten häufig auch eine lexikalische Schicht, die den Fachwortschatz beschreibt, mit dem die konzeptuelle Ebene sich in Texten wiederfindet. Mit der lexikalischen Schicht können bei Suchanwendungen Sprachvariationen (Synonyme, Abkürzungen, unterschiedliche Fachsprachen) oder auch Mehrsprachigkeit unterstützt werden. Die zur Wissensorganisation bekannten Thesauri stellen “leichtgewichtige” Ontologien dar und können in Semantic-Web-Anwendungen genutzt werden.

Nun sind ausdrucksfähige Metadaten und umfangreiche Thesauri in der Umweltinformatik nicht neu. Trotzdem sind semantische Technologien in Umweltanwendungen noch nicht gängige Praxis. Auch Forschungsarbeiten in diesem Bereich befassen sich überwiegend mit der Suche nach Text- oder Multimedia-Dokumenten und weniger mit der Suche nach Daten. Daher untersucht disy praxistaugliche Kombinationen semantischer Technologien mit Cadenza, mit dem Ziel, eine intuitive, begriffsbasierte Suchschnittstelle für komplexe Daten eines Umwelt-Data-Warehouse zu schaffen, die mit maximal 2 Klicks zu Endergebnissen führen kann. Diese Suchschnittstelle wäre somit ein dritter, begriffsbasierter Zugang zu Cadenza-Inhalten, neben der kartenbasierten und der maskenbasierten Recherche. Wir skizzieren zwei aktuelle Prototypen in diesem Kontext, HIPPOLYTOS und KOIOS. [1, 2]

[1] HIPPOLYTOS ist ein Projekt für kleinere und mittlere Unternehmen im Rahmen des BMWi-Forschungsprogramms THESEUS („Neue Technologien für das Internet der Dienste“), in dem disy zusammen mit Fraunhofer IOSB Suchschnittstellen für Geo- und Umweltdaten untersucht. HIPPOLYTOS-Konzepte basieren auf dem KEWA-Projekt SUI /2/ und ergänzen komplementär die Ergebnisse von KEWA-Projekt SUI II /3/. Der KOIOS-Prototyp wurde im Rahmen des THESEUS-Teilprojekts CTC-WP3 am FZI entwickelt und im Rahmen einer Diplomarbeit /4/ auf Cadenza angepasst. Details zur technischen Realisierung von HIPPOLYTOS und KOIOS finden sich in /4/ bzw. /5/.

2. Semantische Suche mit expliziten Metadaten

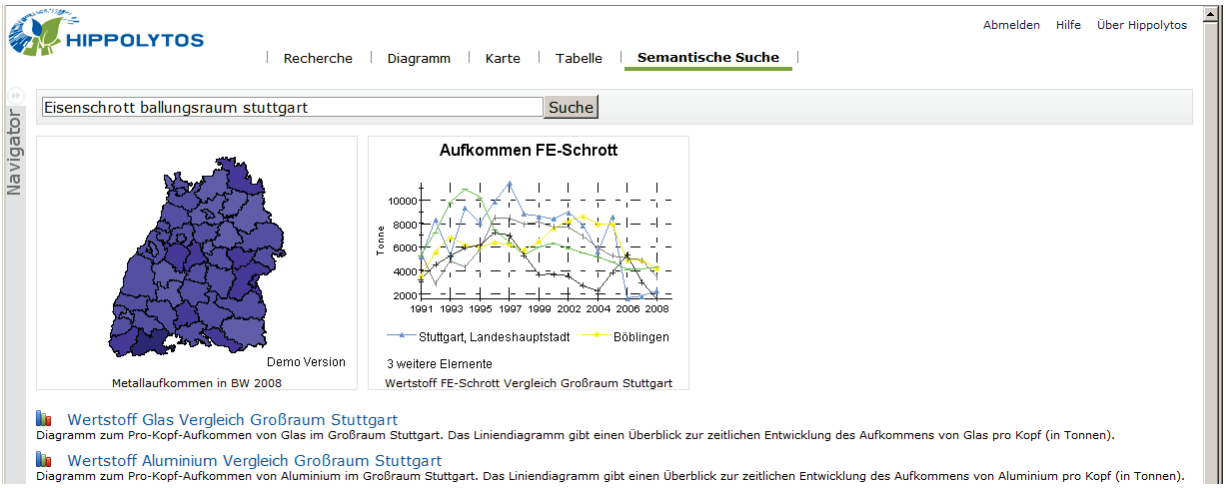

Abbildung 1: Ausschnitt der Ergebnisliste der HIPPOLYTOS-Suche für „Eisenschrott Ballungsraum Stuttgart“

Abbildung 1 zeigt die Ergebnisliste von HIPPOLYTOS zur Anfrage „Eisenschrott Ballungsraum Stuttgart“. Die Ergebnisliste wird immer angeführt von der Live-Vorschau der jeweils höchst rangierenden Cadenza-Selektoren[3], die eine Karten- bzw. eine Diagrammdarstellung liefern. Folgende Schlüsse wurden bei der Anfrage-Auswertung unter anderem gezogen:

- „Eisenschrott“ ist im unterliegenden Repository kein Fachbegriff – aber „Wertstoff FE-Schrott“ ist einer, mit dem Synonym „Eisenschrott“ in der lexikalischen Schicht.

  Daher wird der Selektor „Diagramm: Vergleich Aufkommen Wertstoff FE-Schrott Großraum Stuttgart“ gefunden (in Abb. 1 oben rechts als Vorschau angezeigt).
- Die Begriffshierarchie der Ontologie kennt „Wertstoffe” als Oberbegriff von „Eisenschrott“, ebenso „Metall“ als Oberbegriff von „Eisen (FE)“ und „Abfall“ von „Schrott“.

  Daher wird der Selektor „Metallaufkommen im Abfall, Vergleich der Landkreise“ gefunden (in Abbildung 1 oben links als Vorschau).
- Die Begriffshierarchie enthält weiterhin „Wertstoffanteil Aluminiumschrott“ und „Wertstoffanteil Glas“ als Geschwister-Themen zu „Wertstoffanteil FE-Schrott“.

  Daher können entsprechende Selektoren (mit geringerer Relevanz) ebenfalls sinnvolle Suchergebnisse sein.
- Weiterhin kann in der lexikalischen Schicht der Ontologie repräsentiert sein, dass „Großraum Stuttgart“, „Metropolregion Stuttgart“ und „Ballungsraum Stuttgart“ Synonyme für einen vagen Begriff sein können, der sich räumlich unterschiedlich interpretieren lässt, bspw. als das Stadtgebiet Stuttgart, die engere Region mit dem Stadtbezirk Stuttgart und 5 umliegenden Landkreisen oder auch als geographische Erstreckung in einem gewissen Radius um das Stadtzentrum.

---

[2] Die Daten für die in diesem Beitrag gezeigten Demonstratoren stammen mit freundlicher Genehmigung vom Umweltministerium Baden-Württemberg (UM BW), vom Landesamt für Geoinformation und Landentwicklung (LGL) Baden-Württemberg sowie vom Statistischen Landesamt Baden-Württemberg.

[3] disy Cadenza erlaubt die Definition von Such-, Analyse- und Visualisierungslösungen für raumbezogene Daten. Im Kern steht das Repository-System, das die unterliegenden Datenquellen verwaltet. Zentral ist die Idee des Cadenza-Selektors, einer vordefinierten Anfrage-Schablone für bestimmte Datenquellen, die von Fachexperten für spezifische Analyse-Aufgaben definiert, mit Metadaten beschrieben und abgespeichert werden kann.

Mithilfe von solchem lexikalischen und konzeptuellen Hintergrundwissen können semantisch indizierte Selektoren gefunden werden. Die Abbildung von Anfrage-Konzepten auf semantische Selektor-Metadaten kann sich auf verschiedene Selektor-Aspekte beziehen:

- **Selektor-Thema**: Zum Beispiel könnte es einen Selektor geben, der das Aufkommen an bestimmten *Wertstoffen* [Wertstoff (Eisen, Glas, Aluminium) könnte ein Parameter dieses Selektors sein, der zur Aufrufzeit instanziiert wird] im sortierten *Abfall* einer bestimmten administrativen Region [2. Parameter] in einem bestimmten Zeitraum [3. Parameter] sucht und entsprechend darstellt.

  Bei der Beispiel-Suche nach „Eisenschrott" könnte dieser Selektor, z. B. mit „Wertstoffe; Abfall" indiziert, mit dem o. a. Hintergrundwissen gefunden werden.
- **Wertebereich von Selektor-Parametern:** *„FE" als Synonym von „Eisen" könnte den 1. Parameter des Beispielselektors belegen; „Stuttgart" als Abkürzung für „Stadtkreis Stuttgart" oder „Regierungsbezirk Stuttgart" den 2. Parameter.*
- **Visualisierungs- oder Darstellungsform** der Ergebnisse, z. B. Datenwert(e), Datentabelle, kartenbasierte Darstellung, spezieller Diagrammtyp: Hier können Details der Anfrageformulierung Hinweise auf die erwartete Darstellung geben, z. B. „Vergleich" für ein Balken- oder Kuchendiagramm, „Trend" für ein Liniendiagramm oder „Verteilung" – in einem räumlichen Kontext – für eine Kartendarstellung.

Gefundene Selektoren können dann beim Anklicken mit den entsprechenden Parametern instanziiert und ihre Ergebnisse angezeigt werden.

3. Schemafreie Suche

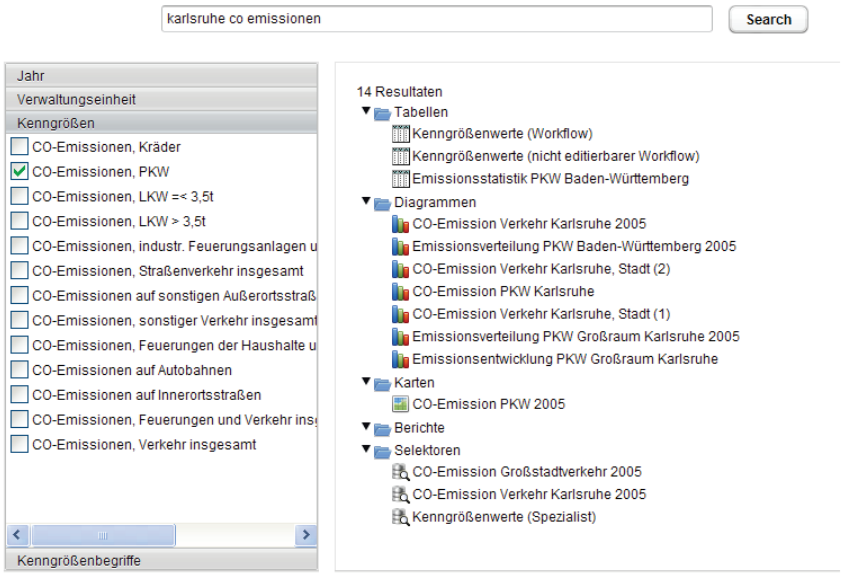

Abbildung 2: Facettierte Suche mit KOIOS für „Karlsruhe CO Emissionen"

KOIOS (s. Abbildung 2) geht grundsätzlich anders vor: Hier wird eine sogenannte schemafreie Suche verwendet, die aus einer Menge von Schlüsselworten eine Reihe möglicher Datenbank-Anfragen erzeugt, die durch diese Schlüsselworte beschrieben werden könnten – ohne dass man dabei das DB-Schema kennen müsste (im Gegensatz dazu steckt das Schema-Wissen bei HIPPOLYTOS in den Selektoren).

Dazu werden die konkreten DB-Inhalte statistisch ausgewertet und es wird ein probabilistisches Modell erstellt, welche Werte wie häufig in welchen Attributen des relationalen Schemas auftauchen. Werden dann bestimmte DB-Werte als Schlüsselworte einer Suche verwendet, kann man mit heuristischen Methoden Hypothesen erzeugen, welche SQL-Anfrage auf dem Schema gemeint sein könnte. Innerhalb von disy Cadenza lassen sich dann wiederum diejenigen Selektoren finden, die diesen Anfragen am nächsten kommen. Da bei komplexen Schemata und häufig auftauchenden Werten i. A. viele Interpretationen möglich sind, werden einerseits Ranking-Methoden zur Sortierung wichtig, andererseits wurde GUI-seitig eine facettierte Suche /6/ realisiert (s. Abbildung 2). Hierbei werden die verschiedenen Dimensionen, hinsichtlich derer sich verschiedene Ergebnismöglichkeiten unterscheiden (in Abbildung 2 links die Selektorparameter Jahr, Verwaltungseinheit, Kenngröße), jeweils mit ihren möglichen Ausprägungen angeboten; selektieren von gewünschten Werten führt dann direkt rechts zur Aktualisierung der Liste möglicher Suchergebnisse.

4. Zusammenfassung

Wir haben die Funktionalität zweier Ansätze skizziert, die eine Schlüsselwortsuche in Cadenza realisieren. Beides sind Prototypen und benötigen noch „Feinschliff“ in der Implementierung sowie weitere Experimente zu Ergebnisgüte und Benutzerfreundlichkeit. Sie zeigen aber, dass Direktzugriffe auf aufbereitete Cadenza-Inhalte prinzipiell machbar sind. Die Ansätze haben komplementäre Eigenschaften und Stärken (bspw. baut HIPPOLYTOS ausschließlich auf Metadaten, KOIOS nutzt diese gar nicht; HIPPOLYTOS kann externe Ontologien nutzen, KOIOS braucht sie nicht; KOIOS interpretiert die Werteverteilung der realen Daten intelligent, versagt bei nicht in der DB auftauchenden Suchbegriffen, was bei HIPPOLYTOS gleichgültig ist). Weitergehende Kombinationen sind daher vielversprechend.

5. Literatur

/1/ Domingue, R.; Fensel, D.; Hendler, J. A.; Hrsg. (2011): Handbook of Semantic Technologies, Springer-Verlag, Berlin, Heidelberg.

/2/ Abecker, A. et al. (2009): SUI – Ein Demonstrator zur semantischen Suche im Umweltportal Baden-Württemberg. Mayer-Föll, R. et al.; Hrsg.: Kooperative Entwicklung wirtschaftlicher Anwendungen für Umwelt, Verkehr und benachbarte Bereiche in neuen Verwaltungsstrukturen, Phase IV 2008/09, Forschungszentrum Karlsruhe, Wissenschaftliche Berichte, FZKA 7500, S. 157-166.

/3/ Bügel. U. et al. (2010): SUI II – Weiterentwicklung der diensteorientierten Infrastruktur des Umweltinformationssystems Baden-Württemberg für die semantische Suche nach Umweltinformationen. In: Mayer-Föll, R. et al.; Hrsg.: Kooperative Entwicklung wirtschaftlicher Anwendungen für Umwelt, Verkehr und benachbarte Bereiche in neuen Verwaltungsstrukturen, Phase V 2009/10, Karlsruher Institut für Technologie, KIT Science Reports, FZKA 7544, S. 43-50.

/4/ Nedkov, R. (2011): Schlüsselwortsuche über relationalen Datenbanken, Diplomarbeit, Karlsruher Institut für Technologie.

/5/ Abecker, A. et al. (2011): Enabling User-friendly Query Interfaces for Environmental Geodata through Semantic Technologies. In: Schwering, A. et al.; Hrsg.: GEOINFORMATIK 2011 – GEOCHANGE, Akademische Verlagsgesellschaft Aka, Heidelberg.

/6/ Tunkelang, D. (2009): Faceted Search (Synthesis Lectures on Information Concepts, Retrieval and Services), Morgan & Claypool.