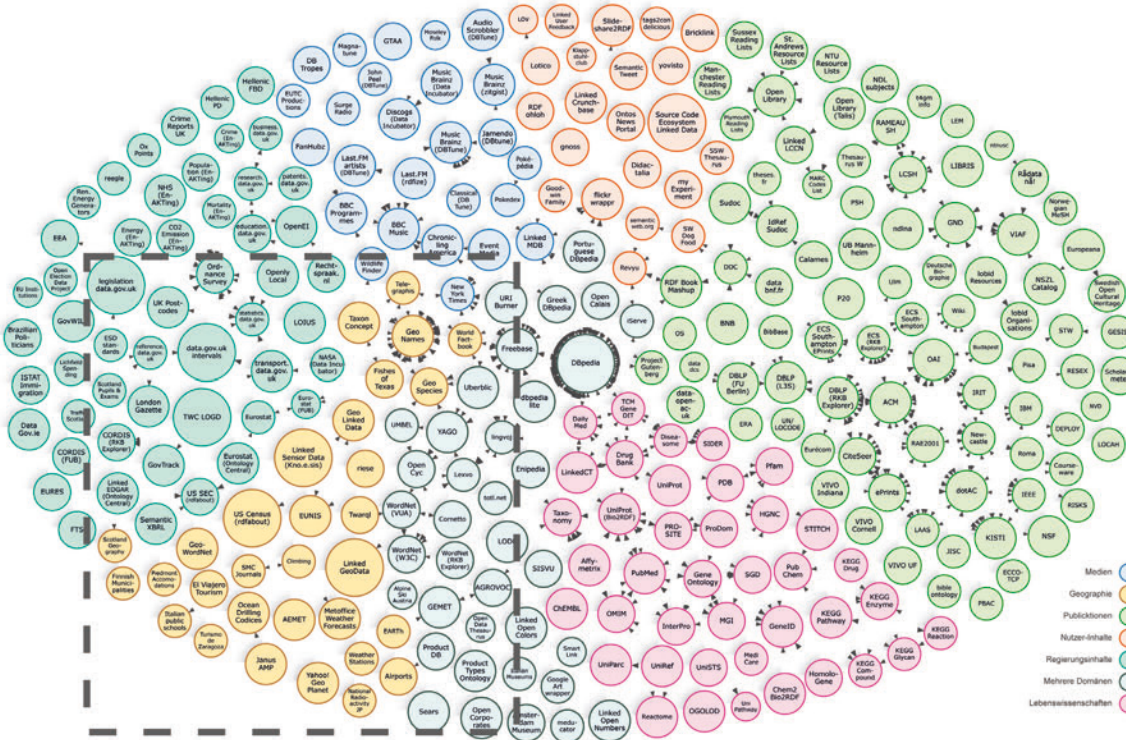




Öffentliche Wissensbasen im Web: Geodaten spielen eine herausragende Rolle, sind aber nur unzureichend verknüpft. Das will GeoKnow ändern.

GEO-ANWENDUNGEN IM DATENWEB

Räumliche Daten und Informationen sind heute im Internet weitreichend verfügbar. An deren Verknüpfung mit nutzerrelevanten Informationen hapert es jedoch zumeist. GeoKnow setzt hier an.

Neben den expliziten räumlichen Daten haben auch viele Webanwendungen einen geographischen Bezug. Das obige Bild zeigt öffentliche Wissensbasen im Internet sowie deren Verknüpfungen. Dabei wird die starke Rolle von Geodaten – gelb markiert und eingerahmt – deutlich.

Trotz umfangreicher Datenbanken werden viele dieser sogenannten Geo-Daten derzeit nur unzureichend verknüpft. Kartendienste wie Google Maps, Yahoo Maps und Microsoft Live Maps zeigen Standorte von Geschäften und deren Kundenbewertungen an. Momentan sind mögliche

Suchanfragen an diese Kartendienste aber noch sehr begrenzt, da es bisher nicht möglich ist, geographische Standorte mit nutzerrelevanten Informationen, wie den Öffnungszeiten eines Geschäftes, zu verknüpfen. Die erforderlichen Informationen, um solche Fragen zu beantworten, sind zwar vorhanden, aber auf eine Vielzahl von Informationsquellen und Datenbanken verteilt, die nicht miteinander verknüpft sind. Das GeoKnow-Projekt mit der Projektkoordination an der Universität Leipzig setzt sich zum Ziel, diese Situation zu verbessern, und soll in diesem Artikel kurz beschrieben werden.

Geodaten sind essenziell für geschäftskritische Entscheidungen und wichtig für die Verwaltung von lokalen, nationalen und globalen Ressourcen. Eine Karte über hochwassergefährdete Gebiete, Höhe bisheriger Hochwasser oder vorhandene Schutzmaßnahmen kann beispielsweise eine wichtige Rolle für die Bebauungsplanung spielen oder im Fall einer Überschwemmung bei der Koordinierung der Rettungsmaßnahmen helfen. Die Generierung entsprechender Daten ist aufwendig und teuer. Oft

sind die gewünschten Daten zwar schon vorhanden, jedoch nicht integriert oder nicht frei verfügbar, in verschiedenen untereinander inkompatiblen Formaten gespeichert oder nicht umfassend genug. Für eine Weiterverwendung der Daten und als Voraussetzung für die Verknüpfung mit anderen Datensätzen, ist eine Bereitstellung der Daten in standardisierten Formaten notwendig. Ziel hierbei ist, aus den Datensätzen einen Mehrwert zu generieren, der über den ursprünglich geplanten Einsatzzweck der einzelnen Datensätze hinausgeht.

In unserem Beispiel könnte das Höhenprofil eines Region mit einer Flusskarte verknüpft werden, um daraus flutgefährdete Gebiete zu errechnen. Zu diesem Zweck wurden die sogenannten Geodateninfrastrukturen (GDIs) geschaffen sowie die gemeinnützige Organisation Open Geospatial Consortium (OGC) gegründet, die offene Standards für Geodaten festlegt. Durch diese Bemühungen können Geodaten bereits heute zwischen unterschiedlichen Geoinformationssystemen (GIS) ausgetauscht werden. Diese Systeme sind allerdings in der Regel isoliert, nicht im Web vertreten und nur für Experten zugänglich. Eine mögli-

che Lösung bieten hier Technologien des Datenwebs beziehungsweise des semantischen Webs. Dabei bekommt jede Resource, etwa der Stadt Leipzig, einen eindeutigen Identifikator (eine sogenannte URI – Uniform Resource Identifier), mit der diese global identifiziert werden kann. Auf diese Weise können Informationen zu Leipzig aus verschiedenen Wissensbasen, wie touristische Informationen aus einer Wissensbasis und Reiseinformationen aus einer anderen Wissensbasis, miteinander integriert werden. So wird auch das Problem der Mehrdeutigkeit in der natürlichen Sprache gelöst. Mit der Zeichenkette „Leipzig“ könnten zum Beispiel verschiedene Städte in Nordamerika gemeint sein, diverse Schiffe oder zum Beispiel der Landkreis (und nicht die Stadt) Leipzig. Durch den Ersatz der Zeichenketten für diese Objekte durch URIs kann zwischen diesen Bedeutungen unterschieden werden. Auf URIs baut der Standard RDF (Resource Description Framework) auf. Mithilfe von RDF lassen sich Fakten ausdrücken, die sowohl Menschen als auch Maschinen verstehen können. Dieses kombinierte Wissen ergibt eine globale Graphstruktur (z.B. Abb. 1), die Millionen von Fakten enthalten kann. Dieses Netz nennt man Linked Open Data (LOD) Cloud. Aktuell ist dieses Wissen jedoch nur unzureichend an die Bedürfnisse im Umgang mit Geodaten angepasst.

Daher haben wir uns im GeoKnow-Projekt, das zum 1. Dezember 2012 begann, zum Ziel gesetzt, die Vernetzung solcher Datenbanken und Informationsquellen zu vereinfachen und die Bearbeitung, Visualisierung und Verknüpfung von Geodaten zu verbessern. Um diese Ziele zu verwirklichen, plant das Forschungsprojekt neben der Optimierung von bereits bestehenden Werkzeugen auch die Entwicklung neuer Open-Source Werkzeuge. Diese sollen Webnutzern, Unternehmen und Organisationen helfen, strukturierte Geodaten im Web zu finden und zu verwerten. GeoKnow richtet sich an diverse Endnutzer. Derzeit werden die

Technologien vor allem mit Unternehmen der Reisebranche und des Lieferkettenmanagement-Sektors erprobt. Nutzer aus der Reisebranche profitieren beispielsweise von besser vernetzten Hintergrundinformationen, erweiterten Inhalten und verbesserten Geo-Suchfunktionen, die zum Beispiel für die Urlaubsbuchung verwendet werden können. Diesen vielfältigen Herausforderungen der Integration und Verwertung von Geodaten im Web widmen sich sechs Partner aus Industrie und Forschung. Dazu gehören neben dem Research and Innovation Center aus Griechenland (Athena RC), Open Link Software aus Großbritannien, Unister und Brox aus Deutschland sowie die Firma Ontos aus der Schweiz. Koordiniert wird das auf drei Jahre angelegte EU-Projekt durch Dr. Jens Lehmann innerhalb der Forschungsgruppe Agile Knowledge Engineering and Semantic Web (AKSW) des Instituts für Angewandte Informatik (InfAI) an der Universität Leipzig.

OFFENE GEODATEN IM WEB

Momentan gibt es drei Hauptquellen für offene Geodaten im Web: Geodateninfrastrukturen (GDIs), offene Datenkataloge und Crowdsourcing-Initiativen.

Geodateninfrastrukturen (GDIs) sind komplexe Forschungsnetzwerke, die dafür sorgen, dass Geodaten nicht isoliert sind, sondern dass alle Beteiligten auf sie zugreifen können. GDIs bestehen sowohl aus technischen Strukturen, wie Servern und Datenbanken, als auch aus organisatorischen Strukturen, Verfahrensweisen und Standards zur effizienten Suche und Nutzung geographischer Daten im Web. Die Weiterentwicklung von GDIs wird vor allem von internationalen Organisationen wie ISO/TC 211 Geographic information/Geomatics, dem Open Geospatial Consortium OGC und dem World Wide Web Consortium W3C geleitet. Die Europäische Union baut eine eigene GDI auf, die ESDI. Des Weiteren gibt es die

INSPIRE-Richtlinie, die den Standards des OGC folgt, um die Interoperabilität der Geodatenätze der verschiedenen europäischen Mitgliedsstaaten zu verbessern. Damit stellt die Richtlinie die legalen und technischen Voraussetzungen bereit, damit die GDIs aller Mitgliedsstaaten miteinander kompatibel und grenzübergreifend nutzbar sind.

Offene Datenkataloge stellen Geodaten, ähnlich wie eine Bibliothek, geordnet zur Verfügung, beispielsweise nach Größe und Abdeckung geordnet, sodass sie einfacher zu finden sind. Sie stellen Geodaten bereit, indem sie (a) existierende GDIs als Basis verwenden oder (b) verfügbare Geodaten als Dateien bereitstellen. Diese Dateien enthalten sowohl strukturierte als auch frei verfügbare Daten.

Beim **Crowdsourcing** werden die Daten von den Benutzern selbst angelegt. Dadurch ist die Anzahl der „Mitarbeiter“, die zusammen schnell eine sehr große Menge an Geodaten erfassen können, potenziell sehr groß. Oft sind dann Informationen verfügbar, die offiziell noch nicht erfasst sind, wie beispielsweise kleine Feld- und Waldwege. Die wichtigsten, durch Crowdsourcing entstandenen geographischen Wissensdatenbanken sind OpenStreetMap, GeoNames und Wikipedia. GeoNames stellt grundlegende geographische Daten bereit, wie Längen- und Breitengrad, Höhe über dem Meeresspiegel, Bevölkerungszahl, Verwaltungseinheiten und Postleitzahlen. Diese Daten sind als Textdateien verfügbar und durch viele Webservices zugänglich. Durch die Bereitstellung einer Vielzahl von Daten ist OpenStreetMap (OSM) eine wichtige Plattform für das Verknüpfen, Durchsuchen und Visualisieren von Geodaten im Web. Die Daten sind in verschiedenen Formaten verfügbar und können in eine Datenbank importiert werden. Durch OSM werden zudem Webservices, wie Reverse Geocoding (Herauffinden der Adresse an einem bestimmten Punkt) bereitgestellt.



Making the web an
exploratory place
for geospatial data

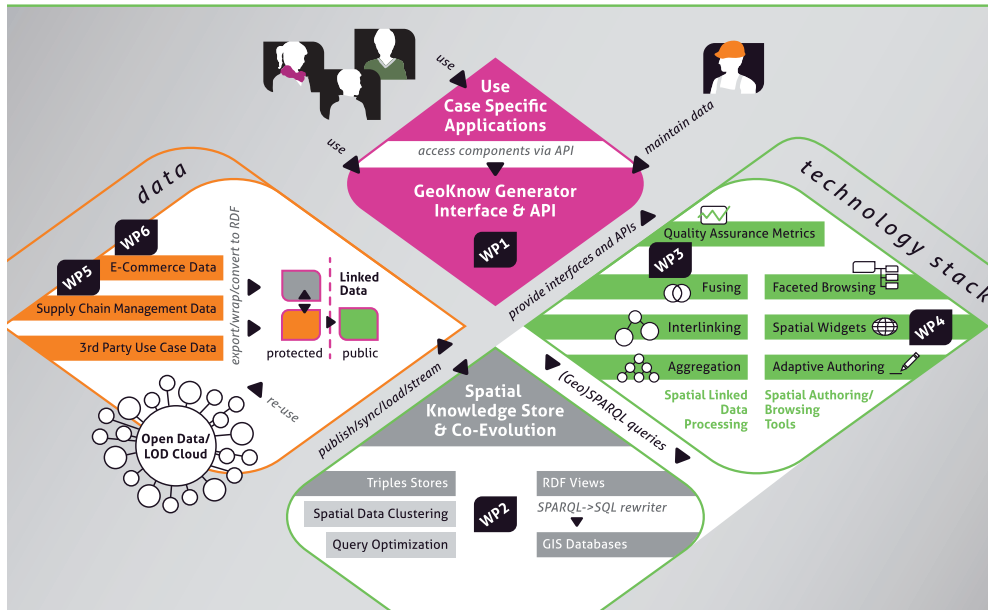


Abbildung 1: EU, Forscher und Industrie arbeiten zusammen im GeoKnow-Projekt, um die Vernetzung, Integration und Visualisierung von Geodaten im Web zu verbessern.

SEMANTISCHE TECHNOLOGIEN FÜR GEODATEN IM WEB

Semantische Technologien für das Management von den soeben beschriebenen Geodaten ermöglichen eine intelligentere Nutzung der Daten. Zum Beispiel lassen sich die Daten mithilfe von sogenannten Ontologien strukturieren. Diese Ontologien beschreiben die verwendeten Wissensstrukturen und ermöglichen deren einheitliche Verwendung – selbst wenn die Daten von vollkommen unterschiedlichen Organisationen an verschiedenen Orten erstellt werden. Leider bieten aktuelle GIS-Werkzeuge noch eingeschränkte Unterstützung für diese ontologischen Strukturen. Deshalb gibt es in diesem Bereich noch zahlreiche Verbesserungsmöglichkeiten, die innerhalb von GeoKnow erforscht werden sollen.

Frühere Ansätze in dieser Richtung umfassen das Basic Geo Vocabulary der W3C, welches die Repräsentation von Punkten im WGS84-Referenzsystem erlaubt und GeoRSS, das die Darstellung komplexer räumlicher Objekte, wie Lini-

en, Rechtecke und Polygone erlaubt. Ein weiterer Ansatz ist GeoOWL, das ein flexibles Modell für räumliche Konzepte bereitstellt. Weiterhin kann die topologische Modellierung von geometrischen Formen in RDF auch mit der sogenannten NeoGeo Geometry Ontology durchgeführt werden. All diese Ontologien unterstützen allerdings ausschließlich WGS84 und bieten nur begrenzte Unterstützung für die räumlichen Operationen (z.B. nach RCC-8), die unter realen Bedingungen in GIS auftreten.

GeoSPARQL ist ein erfolgversprechender Standard des W3C, um RDF-Geodaten auszudrücken und abzufragen mit dem Ziel, das Einfügen und Abfragen von Geodaten zu standardisieren. GeoSPARQL erlaubt begrenztes Reasoning, also das Schlussfolgern von Tatsachen aus existierenden Informationen. Ein einfaches Beispiel: Auf der einen Seite werden die Standorte aller Firmen erfasst, die Dachdecker anbieten. Eine andere Definition ordnet Dachdecker der Gruppe Handwerksbetriebe zu. Durch diese Schlussfolgerungen kann der Nutzer direkt alle Handwerksbetriebe, unter

anderem eben auch alle Dachdeckerfirmen, auf einer Karte anzeigen.

GeoSPARQL orientiert sich weiterhin streng an den existierenden Standards des OGC für Geodaten und erleichtert damit das schnelle Zugreifen auf Informationen innerhalb einer Datenbank.

GEOKNOW-ARCHITEKTUR

Motiviert ist die Arbeit in GeoKnow durch vorhergehende Forschungsarbeit am Projekt LinkedGeoData (LGD), das OpenStreetMap-Daten in Form einer RDF-Wissensbasis erstellt. Durch diese Daten konnten Verknüpfungen zu anderen Wissensbasen in der LOD-Cloud hergestellt werden, beispielsweise zu GeoNames und DBpedia (Wikipedia in RDF-Form). LGD war dafür bestimmt, einfache Geodaten zu generieren und zu aggregieren. In dieser Zeit wurden mehrere Herausforderungen identifiziert, wie die Skalierbarkeit von Abfragen auf Geodaten, die Abfrageperformanz, Datenmodellierung, flexible Transformation von Spezialdaten und Datenoperationen wie das Routing. Es stellte sich heraus, dass Geodaten im Web neue Möglichkeiten für Anwendungen im Management und in der Entscheidungsfindung eröffnen. Die grundlegende Struktur, die es GeoKnow ermöglicht, diese Anwendungen zu nutzen (siehe Grafik Architektur), lässt sich in vier Gebiete zerlegen. In diesen Bereichen stehen jeweils unterschiedliche Aspekte im Mittelpunkt: die Behandlung der Daten, die verwendeten Technologien, die Wissensbasen sowie die Anwendungen und Benutzerschnittstellen.

Bei der Verwertung der Daten (links in der Grafik) werden zwei Anwendungsfälle betrachtet: E-Commerce und Supply Chain Management. Im E-Commerce-Anwendungsfall sollen in Zusammenarbeit mit der Firma Unister, die viele populäre Webportale betreibt, Verbesserungen im Tourismusbereich erreicht werden. Insbesondere soll die Suche nach Reiseorten und Empfehlungen für Nutzer verbessert werden.

Quelle: GeoKnow

Im Bereich Supply Chain Management sollen die Versorgungsketten bei einem großen deutschen Automobilhersteller transparenter und effizienter gestaltet werden. Semantische Technologien erlauben auch hier, wie oben beschrieben, die Integration von Daten vieler verschiedener Firmen in einem komplexen Netzwerk von Zulieferern in der Automobilindustrie.

Im Bereich des Speicherns und Abfragens von Daten (unterer Bereich der Grafik) geht es hauptsächlich um das Einbinden existierender GIS-Systeme, sowie die Verbesserung der Effizienz bei der Abfrage von Wissen, das Geoinformationen mit strukturiertem Wissen kombiniert.

Darauf aufbauend werden Verfahren zur Visualisierung und Integration der Daten (rechter Bereich der Grafik) entwickelt. Die Datenintegration erfolgt dabei halbautomatisch. Konkret können dabei mit niedrigem Aufwand von Experten selbst große Datenbestände mit Millionen von Fakten integriert werden. In GeoKnow werden weiterhin effiziente Qualitätssicherungsmaßnahmen für semantische Geodaten untersucht, um den Ansprüchen vom Einsatz in Unternehmen gerecht zu werden.

Aufbauend auf diesen Komponenten wird eine integrierte Plattform entwickelt (oberer Bereich der Grafik). Diese basiert auf Debian-Paketen, die auch in Linux-Distributionen die Beziehungen zwischen Softwarekomponenten regeln. Außerdem erlauben diese Pakete die leichte Installation im Serverbereich. Abgesehen davon werden auch Programmierschnittstellen entwickelt, die Entwicklern erlauben selbst semantische Geoanwendungen zu entwickeln, unter anderem unter Nutzung der GeoKnow-Komponenten.

BEISPIEL

Verknüpfte Geodaten spielen eine zentrale Rolle in einer Vielzahl von Endnutzer-Szenarien. Zum Beispiel enthalten Wissensbasen wie Linked-GeoData Informationen zur Lage von

Schulen in Oxford. Andere Wissensbasen stellen Geodaten zur Lage von Mietobjekten in der gleichen Stadt zur Verfügung. Um Mietobjekte einer bestimmten Größe in der Nähe von Privatschulen schneller zu finden, mussten Endnutzer bisher Daten zur Lage von Mietobjekten in Oxford aus diversen Webseiten zusammentragen. Zusätzlich musste die Lage von entsprechenden Schulen, die ihren Kriterien entsprechen, gespeichert werden. Anschließend galt es, die Lage der Schulen und der Mietobjekte manuell miteinander zu vergleichen, um die gesuchten Mietobjekte zu finden. Durch die Verknüpfen der Geodaten aus mehreren Quellen und die Entwicklung von Endnutzeranwendungen, die auf diesen Daten basieren, wird das GeoKnow-Projekt Technologien und Datensätze zur Entwicklung von Question-Answering-Systemen zur Verfügung stellen, die es erlauben werden, solche komplexe Fragen vollautomatisch – also ohne manuelles Zusammentragen von Lageinformationen – zu beantworten. Erste Algorithmen in diese Richtung wurden bereits im Question Answering System TBSL (<http://autosparql-tbsl.dl-learner.org/>) implementiert. Hier werden die Daten aus LinkedGeoData mit denen des Datenextraktionssystem OXPath mittels des Frameworks LIMES miteinander verknüpft. Mithilfe dieses Systems können nun Anfragen wie „Gib mir Häuser mit Balkon in der Nähe von Schulen“ gestellt und beantwortet werden. Solche Abfragen, die die Informationen mehrerer Makler zusammentragen, sind mit herkömmlichen Technologien bisher nicht möglich gewesen.

AUSBLICK

Unsere Vision ist es, Geodaten im Datenweb zu veröffentlichen und damit das Web zu einem Ort zu machen, in dem Geodaten ähnlich zu anderen Webinhalten veröffentlicht, abgefragt, verknüpft und für Schlussfolgerungen genutzt werden können. Damit rei-

chen Geodaten über die syntaktische Interoperabilität hinaus und erlauben semantische Interoperabilität und Systeme, die räumliches Schlussfolgern im Web ermöglichen. Semantische Technologien werden damit nicht nur um Geodaten erweitert, sondern auch befähigt, Antworten auf Fragen zu geben, die mit isolierten Quellen von Geodaten bisher nicht beantwortet werden konnten. Insbesondere liegt der Beitrag des Projekts in der effizienten Abfrage von RDF-Geodaten, der Verknüpfung und Integration von Daten, deren Visualisierung und letztendlich der Neu- und Weiterentwicklung von zahlreichen Werkzeugen für die Geo-Gemeinschaft. ◀

AUTOREN UND KONTAKT:

Jens Lehmann

GeoKnow Koordinator
Universität Leipzig, Deutschland
lehmann@informatik.uni-leipzig.de

Konrad Höffner

Wissenschaftlicher Mitarbeiter
Universität Leipzig, Deutschland

Sandra Prätor

GeoKnow Management
Universität Leipzig, Deutschland

Stephanie Lehmann

Wissenschaftliche Mitarbeiterin
TU Dresden

Axel Ngonga

Forschungsgruppenleiter
Universität Leipzig, Deutschland

Alejandra Garcia-Rojas

GeoKnow Arbeitspaketsleitung
Ontos, Schweiz

Spiros Athanasiou

Wissenschaftlicher Mitarbeiter
Athena RC, Griechenland