Verwendung von Wiktionary zur Verbesserung der Discovery in geoplatform.at

Manfred MITTLBÖCK und Bernhard VOCKNER
Research Studios Austria – iSPACE, Salzburg · manfred.mittlboeck@researchstudio.at

Zusammenfassung

Geoportale dienen als Zugangspunkte zu Geodateninfrastrukturen (GDI), um (geographische) Information zu finden, zu evaluieren und zu verwerten. Eine große Herausforderung in Bezug auf das Auffinden und Entdecken von Information ergibt sich aufgrund der semantischen Heterogenität der Dokumentationen. Letztere entsteht dadurch, dass in der Metadatenerfassung in verschiedenen Anwendungsdomänen unterschiedliche (fachspezifische) Begrifflichkeiten bzw. fremdsprachige Begriffe zur Beschreibung von Ressourcen Anwendung finden. Um mit der semantischen Heterogenität besser umgehen zu können, schlagen wir eine semantische Erweiterung von Metadaten mithilfe von Ontologien und Thesauri vor. Dazu haben wir Werkzeuge entwickelt, die eine automatische Extraktion von Schlüsselwörtern, deren Synonymen, Ober- und Unterbegriffen sowie Übersetzungen in andere Sprachen aus Ontologien und Thesauri ermöglicht. Die daraus extrahierten Termini reichern Metadaten mit zusätzlichen Inhalten an und können damit die Discovery Experience bei der Suche in Geoportallösungen erheblich verbessern.

1 Einleitung

Geoportale sind zentrale Zugriffspunkte auf räumliche und vermehrt auch auf nichträumliche Informationen. Ein wesentlicher Nachteil bestehender Geoportalimplementierungen ist sehr häufig, dass diese aus Sicht der Anwender nur mangelhaft mit Suchanfragen umgehen können, d. h. oft unzureichende Suchergebnisse liefern. Dies tritt vor allem in Fällen auf, in denen der vom Nutzer eingegebene Suchbegriff nur bedingt mit den dokumentierten Begriffen der gesuchten Datensätze übereinstimmt (SMITS & FRIIS-CHRISTENSEN 2007). Dieses Phänomen, das auch als ‚Semantische Heterogenität' bezeichnet wird, tritt vor allem dann gehäuft auf, wenn Vertreter unterschiedlicher Domänen in einer gemeinsamen anwendungsagnostischen Geoportal Zusammenschau nach Daten suchen (ANDRADE & BAPTISTA 2011). Zusätzlich können Schwierigkeiten auftreten, wenn sich die Sprache der Suchanfrage auf Seite Suchsprache des Geoportalnutzers von jener des Datenbereitstellers unterscheidet (NOWAK et al. 2005, OSTLÄNDER & LUTZ 2008). Beispielsweise sind in Österreich Ressourcen Großteils in Deutsch dokumentiert, obwohl diese oftmals auch für anderssprachige Nutzer von Interesse sind. Daher ist ein multilingualer Suchansatz mit automatischer Übersetzung der Suchresultate in die eigene Sprache wünschenswert.

Um diese Anforderungen besser erfüllen zu können und trotz unterschiedlicher Verwendung von Suchbegriffen und Sprachen bessere Ergebnisse zu liefern, präsentieren wir die semantische Erweiterung von Metadaten mithilfe von Ontologien und Thesauri. Diese

Strobl, J., Blaschke, T., Griesebner, G. & Zagel, B. (Hrsg.) (2014): Angewandte Geoinformatik 2014.
© Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach. ISBN 978-3-87907-543-0.
Dieser Beitrag ist ein Open-Access-Beitrag, der unter den Bedingungen und unter den Auflagen der Creative Commons Attribution Lizenz verteilt wird (http://creativecommons.org/licenses/by/3.0/).

werden in der Literatur als Methoden zum effizienten Umgang mit semantischer Heterogenität beschrieben (WACHE et al., 2001 SMITS & FRIIS-CHRISTENSEN 2007; JANOWICZ et al. 2008).

In der prototypischen geoplatform.at (siehe Abb. 1) haben wir exemplarisch Wiktionary als multilingualen Thesaurus verwendet, um Suchresultate basierend auf Synonymen, Toponymen und anderssprachigen Begriffen zu verbessern.

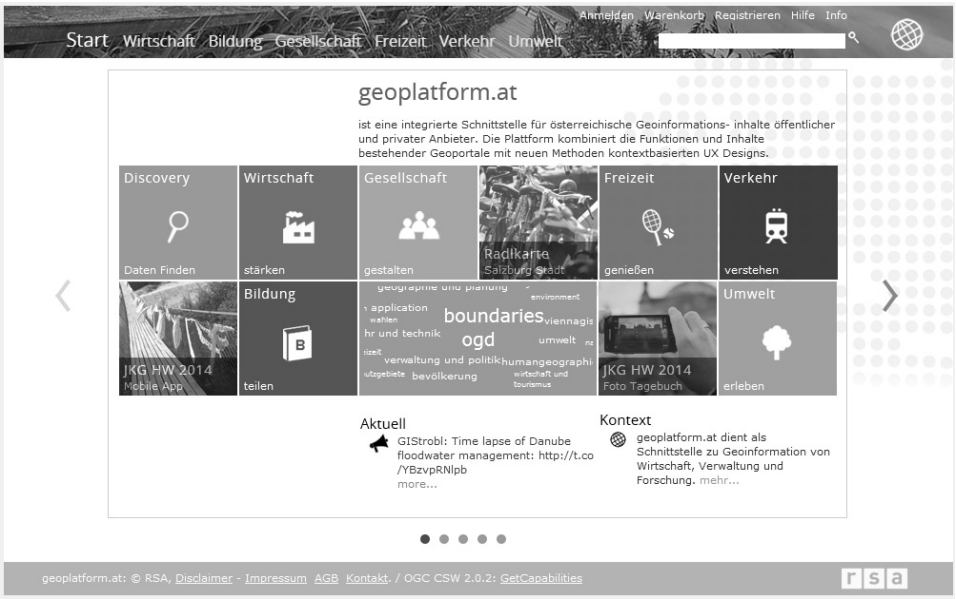

Abb. 1: geoplatform.at: Eine zentrale Plattform um Wissen zu teilen

Im Moment umfasst Wiktionary 363.500 deutschsprachige Einträge, deren qualitative Korrektheit, ähnlich wie bei Wikipedia durch eine breite Community definiert wird. Aufgrund der Begriffsvielfalt und der laufenden Aktualisierung sowie personalisierter Qualitätssicherung haben wir in diesem Ansatz Wiktionary gegenüber Fachthesauri bevorzugt für unsere semantische Erweiterung herangezogen. Dies geht einher mit der Tatsache, dass wir geoplatform.at möglichst benutzerfreundlich gestalten wollten.

2 Geoplatform.at

Als Grundlage für unseren Ansatz der semantischen Erweiterung von Metadaten stehen uns Dateninhalte von der ‚Cooperation OGD Österreich', den INSPIRE Österreich Geoportalinhalten über das ‚Bundesamt für Eich- und Vermessungswesen' (BEV) und dem ‚Land-, forst- und wasserwirtschaftliches Rechenzentrum' (LfRZ)', ‚Geoland.at' und Portalen der einzelnen Bundesländer (z. B. ‚GIS-Steiermark') zur Verfügung. Durch die semantische Verknüpfung dieser Inhalte stellt geoplatform.at eine breite domänenoffene Informationsbasis an raumbezogenen Ressourcen aus Österreich bereit.

3 Semantische Erweiterung von Metadaten in geoplatform.at

Unser Prozessmodell zur semantischen Erweiterung von Metadaten beruht auf Ontologien, Thesauri und SKOS (Simple Knowledge Organization Systems). In geoplatform.at werden die Metadatendokumente aus den in Kapitel 2 angeführten Metadatenkatalogen semantisch erweitert. Das dahinterliegende technische System basiert auf Geoportal Server 1.2.4 (Open Source). Zur Extraktion der Schlüsselwörter aus den Metadatendokumenten verwenden wir Python 2.7 basierte Scripte.

Neben den Werkzeugen zur Extraktion der Schlüsselwörter aus den Metadatendokumenten haben wir noch weitere Tools zum Finden von Synonymen, Ober- und Unterbegriffen sowie deren automatisierter Übersetzungen entwickelt, mit denen die bestehenden Metadatensätze ergänzt werden.

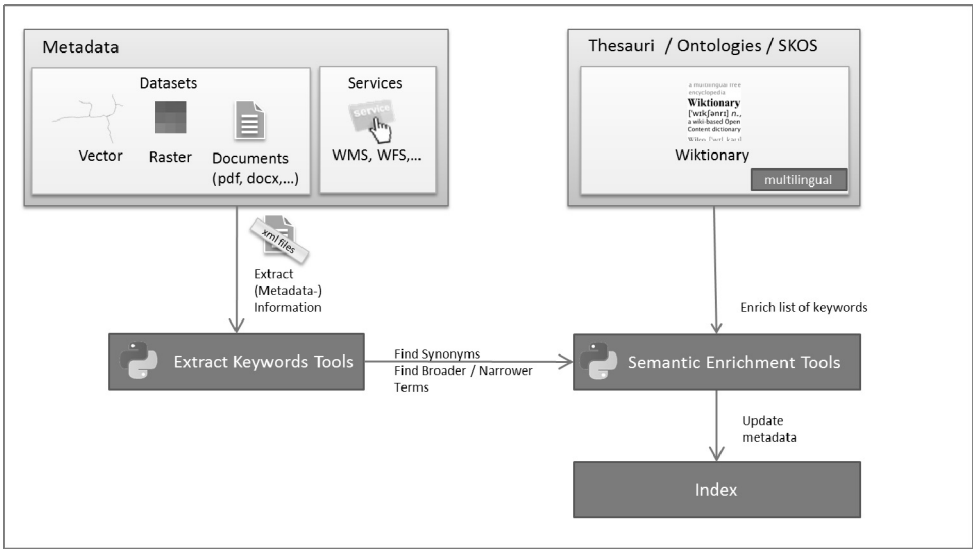

Abb. 2: Schema der semantischen Erweiterung der Discovery in geoplatform.at

Abb. 2 zeigt den Arbeitsablauf für die semantische Erweiterung von geoplatform.at:

In einem ersten Schritt werden Informationen aus bestehenden Metadatensätzen (vorwiegend Schlüsselwörter und Textbaustein-Schlüsselwärter) aus geoplatform.at extrahiert. Angewandt wird dieses Werkzeug für Metadatensätzen die für Vektor- und Rasterdaten, Dokumenten (pdfs, docx) sowie Geodiensten (WMS, WFS) zur Verfügung stehen.

Diese extrahierten Schlüsselwörter dienen als Baustein für die Extraktion von Synonymen sowie generellere (‚broader terms‘) oder spezifischere Begriffe (‚narrower terms‘) aus Wiktionary. Diese neuen Metadateninhalte werden in den bestehenden Metadaten in separaten Sektionen als Schlüsselwörter ergänzt. Diese Erweiterungen dienen dazu, die Metadaten selbst effizienter auffindbar zu machen. Vor allem in Hinblick auf verteilte räumlich-semantische Suchanfragen mit GeoSPARQL (Geographic Query Language for RDF) (OGC

2012) über verschiedene Metadatenkataloge hinweg, ist diese Erweiterung der Metadaten, einer on-the-fly Zuordnung der Vorzug zu geben.

4 Resultate

Um mit der enormen Ressourcenvielfalt in GDIs umgehen zu können, haben wir effiziente Methoden und Strategien zum Auffinden und Entdecken von Information entwickelt. Diese umfassen den Prozessablauf zur Anreicherung von Metadaten mit Synonymen, Ober- und Unterbegriffen sowie Übersetzungen extrahiert aus allgemeinen Ontologien sowie spezialisierten Thesauri und SKOS (Simple Knowledge Organization System) Komponenten. Für die Umsetzung dieser Anreicherungen haben wir Werkzeuge entwickelt, die nach ON/EN/ISO 19139 organisierte Metadaten in Geo-Katalogen z. B. in geoplatform.at anreichern. Diese Ergänzungen ermöglichen die Verbesserung der User-Experience Suchergebnislisten, wenn sich Suchbegriff und Schlüsselwort voneinander unterscheiden, jedoch synonym verwendet oder verwandt sind (Stichwort: broader, narrower).

Als Beispiel kann z. B. in geoplatform.at nach dem Begriff ‚Cabrio' gesucht werden. Diese Suchanfrage hat vor der semantischen Erweiterung mit Wiktionary keine Suchresultate geliefert. Nach der Erweiterung erscheinen Suchergebnisse die Begriffe wie ‚Kraftfahrzeugbestand' und ‚Straßenverkehrsunfälle' beinhalten auch in der Ergebnisliste (Abb. 3).

Abb. 3: Synonymsuche in geoplatform.at

Die oben stehend beschriebene Vorgehensweise ist auch mit englischen Begriffen durchführbar. Die Suche nach ‚car' liefert dementsprechend die gleichen Suchresultate wie die Suche nach dem Begriff ‚Auto' (Abb. 4). Eine automatische Übersetzung der Suchergebnisse in die Sprache der Suche ermöglicht dem Nutzer, ein Grundverständnis über den Datensatz zu erlangen – auch wenn er diese Sprache selbst nicht beherrscht (Abb. 5).

Abb. 4: Englischsprachige Suche in geoplatform.at

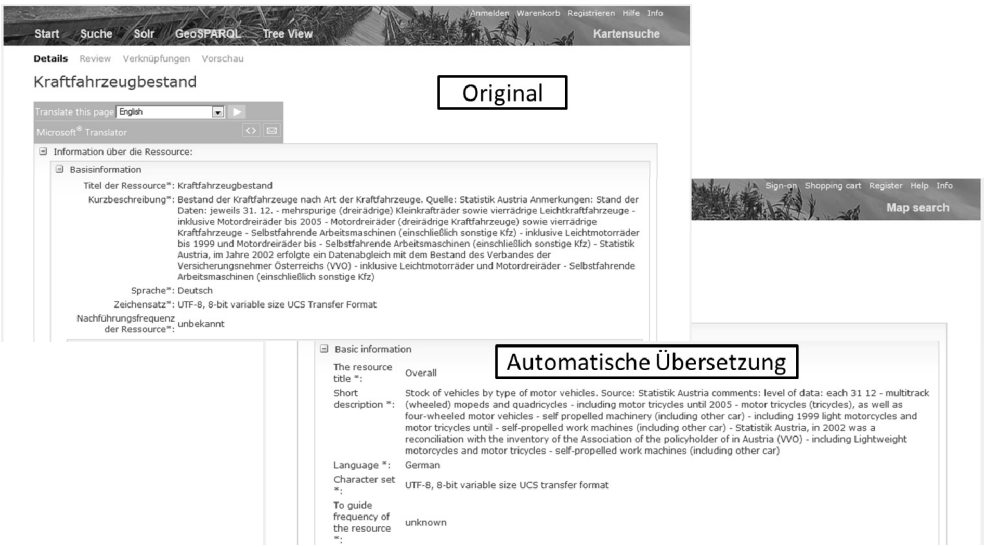

Abb. 5: Automatische Übersetzung der Suchresultate in geoplatform.at

5 Fazit und Ausblick

Wir empfehlen die Integration von semantischen Erweiterungen mit Bezug auf multilinguale Ontologien und Thesauri, veranschaulicht anhand von Wiktionary.org in Geoportal-Lösungen. Zu diesem Zweck haben wir Werkzeuge entwickelt, die eine automatische Extraktion von Schlüsselwörtern sowie deren Synonymen, Ober- und Unterbegriffen sowie Übersetzungen in andere Sprachen ermöglichen und diese Informationen als zusätzliche Information, ergänzt in den Metadaten zu den dokumentierten Ressourcen bereitzustellen. Dieses Set an Werkzeugen und deren Integration in bestehende Geoportallösungen über standardisierte Schnittstellen (OGC CSW − Catalogue Service Web) bilden das Fundament,

die User Experience durch die Verbesserung der Trefferquote bei Suchanfragen entscheidend zu stärken.

Wir wollen darauf hinweisen, dass noch weitere Forschungsarbeit v. a. hinsichtlich der Granulierung unseres Ansatzes notwendig ist. Dies vor dem Hintergrund, dass die automatische Synonymverknüpfung out-of-the box in Hinblick auf die Kontextualisierung schwächen aufzeigt. Dies führt zum Beispiel dazu, dass z. B. bei der Eingabe von ‚Kokain' als Suchbegriff zum Auffinden von Datensätzen zu Drogenkriminalität auch Suchresultate zu Ressourcen über „Schneebedeckung" bzw. ‚Winterdienst' gelistet werden. Dieses lässt sich zwar einfach damit erklären, dass ‚Koks' wie auch ‚Schnee' umgangssprachliche Synonyme für Kokain darstellen. Eine verbesserte Integration und Querabfrage des Such-Kontexts soll hier in Zukunft entsprechende Verbesserungen mit sich bringen. Dazu haben wir geplant, zukünftig die Relevanz von Synonymen sowie deren Ober- und Unterbegriffe zu gewichten und umgangssprachliche Synonyme in unserem Metadatenanreicherungsprozess auszuklammern.

Wir haben unsere Forschungsergebnisse in geoplatform.at validiert. Dieser Forschungsprototyp besteht zurzeit aus mehr als 2.200 deutschsprachigen räumlichen und a-räumlichen Metadatendokumenten aktuell zusammengeführt aus verschiedenen Metadatenkatalogen in Österreich.

Literatur

ANDRADE, F. G. DE & BAPTISTA, C. S. DE (2011), Using Semantic Similarity to Improve Information Discovery in Spatial Data Infrastructures. Journal of Information and Data, 2, 181-194.

JANOWICZ, K., WILKES, M. & LUTZ, M. (2008), Similarity-Based Information Retrieval and Its Role within Spatial Data Infrastructures. In: Proceedings of the 5th International Conference on Geographic Information Science. Springer-Verlag, Park City, UT, USA, Vol. 5266, 151-167.

NOWAK, J., NOGUERAS-ISO, J. & PEEDELL, S. (2005), Issues of multilinguality in creating a European SDI – The perspective for spatial data interoperability. In: Proceedings of the 11th EC GI&GIS Workshop, ESDI Seeting the Framework, Alghero, Italien. 29.06.-01.07.2005.

OGC (2012), OGC GeoSPARQL – A Geographic Query Language for RDF Data. Online 20.04.2014: https://portal.opengeospatial.org/files/?artifact_id=47664.

OSTLÄNDER, N. & LUTZ, M. (2008), INSPIRE-ing GEMET – Enhancing Metadata Creation and Discovery. In: MOELLER, A., PAGE, B., SCHREIBER, M. (Eds.), Environmental Informatics and Industrial Ecology. Proceedings of EnviroInfo 2008 (Lüneburg), 212-214.

SMITS, P. C. & FRIIS-CHRISTENSEN, A. (2007), Resource Discovery in a European Spatial Data Infrastructure. In: IEEE Trans. On Knowl. And Data Eng., 19, 85-95.

WACHE, H., VÖGELE, T., VISSER, U., STUCKENSCHMIDT, H., SCHUSTER, G., NEUMANN, H. & HÜBNER, S. (2001), Ontology-based integration of information – a survey of existing approaches. In: IJCAI-01 Workshop: Ontologies and Information Sharing. Seattle, WA, 04.04.-05.04.2001, 108-117.