Harvesting und Recherche im Geodatenkatalog-DE

Martin HÜBEN

Zusammenfassung

Der Geodatenkatalog-DE ist ein zentraler Dienst, der Katalogdienste bzw. deren Daten zusammenführt, an andere Strukturen abgibt und Recherchen nach Geodaten und Diensten erlaubt. Die Software soll Transparenz über die in Deutschland vorhandenen Geodatenbestände schaffen und die Weiterleitung zu entsprechenden Geodiensten ermöglichen. Das Modul umfasst Software zur Zusammenführung, Konsolidierung und Abgabe von Metadaten, ein entsprechendes Recherchewerkzeug sowie einen INSPIRE-konformen Metadateneditor zur Erfassung eigener Datensätze. Die Zusammenführung der Katalogdienste erfolgt über vorhandene Standard-Schnittstellen (OGC CSW). Ziel dieser Konsolidierung ist es im Wesentlichen, Mehrfacheinträge zu bereinigen und eine zentrale Schnittstelle zu den Metadatenbeständen der GDI-DE zu schaffen. Die Abgabe erfolgt konform zu den Vorgaben der europäischen Initiative zur Schaffung einer Geodateninfrastruktur in der Gemeinschaft (INSPIRE) und der GDI-DE. Für die ersten drei Jahre wird von etwa 50 Katalogdiensten und rund 300.000 Metadatensätzen ausgegangen. Zurzeit stehen im System rund 70.000 Metadatensätze aus ca. 20 Diensten zur Verfügung.

Umgesetzt wird der Geodatenkatalog-DE im Wesentlichen mit den Softwarekomponenten GeoNetwork opensource[1], Mapbender[2] und PostgreSQL[3] als Datenbankinstanz. Hierbei handelt es sich ausschließlich um Open-Source-Software.

1 Ziele des Geodatenkatalog-DE

Ziel des Geodatenkatalog-DE ist die Bereitstellung einer zentralen Recherchemöglichkeit in der GDI-DE über Geoinformationen. Weiterhin ist die Anforderung eines Nutzer (Wirtschaft, Verwaltung, Bürger) einer zentralen Recherche über einen zentralen konsolidierten Geodatenbestand zu erfüllen. Dazu werden die Katalogdienste der verschiedenen Partner der GDI-DE zusammengeführt und somit eine Transparenz über die Geodatenbestände in Deutschland geschaffen. Ebenso soll über den Geodatenkatalog-DE sowohl die Berichtspflicht Deutschlands, welche durch die INSPIRE-Richtlinie 2007/2/EG[4] gefordert wird, unterstützt, als auch die Abgabe von Metadaten an INSPIRE gewährleistet werden.

[1] http://www.geonetwork-opensource.org
[2] http://www.mapbender.org
[3] http://www.postgresql.org
[4] http://inspire.jrc.ec.europa.eu

Strobl, J., Blaschke, T. & Griesebner, G. (Hrsg.) (2011): Angewandte Geoinformatik 2011.

2 Ausgangssituation und Anforderungen

Derzeit liegen die Geoinformationen dezentral in verschiedenen Metadatenkatalogen vor. Dies sind vorwiegend Länderkataloge der Partner der GDI-DE sowie Metadatenkataloge verschiedener Institutionen (PortalU, DLR, etc). Diese bilden teilweise Redundanzen, da viele Metadatensätze von verschiedenen Katalogen vorgehalten werden und über unterschiedliche Dienste recherchierbar sind. Hieraus ergeben sich folgende Anforderungen an das Harvesting und die Recherche im Geodatenkatalog-DE:

- Harvesting:
  1. Dublettenfilter: Es muss ausgeschlossen werden, dass gleiche Metadatensätze, welche in verschiedenen angeschlossenen Metadatenkatalogen vorgehalten werden, im Geodatenkatalog-DE mehrfach vorkommen.
  2. Das Harvesting soll nach erstmaligen Einfügen der Metadaten eines Kataloges nur noch Updates auf den Metadatensätzen durchführen.
- Recherche:
  1. Aufbau eines Index zur performanten Suche.
  2. Gewichtung der Suchergebnisse (Ranking).

Zudem soll der Geodatenkatalog-DE die Katalogschnittstelle über einen Webservice zur Verfügung stellen und über eine standardkonforme Web-Schnittstelle in andere Web-Applikationen integrierbar sein. Auch soll der Geodatenkatalog-DE nachhaltig in Bezug auf Anpassbarkeit sein. Dies ist besonders im Kontext zukünftiger Entwicklungen und Anforderungen, beispielsweise aus INSPIRE, notwendig. Weiterhin ist eine Konformität zu INSPIRE und den Standards von ISO und OGC[5] obligatorisch.

3 Eingesetzte Software

Die Implementierung des Geodatenkatalog-DE erfolgt durch Open-Source-Software. Diese Entscheidung begründet sich hauptsächlich durch die Anforderung an die Nachhaltigkeit der Lösung. Nur so ist gewährleistet, dass zukünftige Anpassungen am Geodatenkatalog-DE herstellerunabhängig erfolgen können, da der Programmiercode offen zugänglich ist und somit Anpassungen von verschiedenen Dienstleistern implementiert werden können. Als Software für den Geodatenkatalog-DE wird GeoNetwork opensource eingesetzt. GeoNetwork ist derzeit die einzige Open-Source-Software im Bereich der Metadatenkataloge, die die Anforderungen an den Geodatenkatalog-DE weitestgehend erfüllt.

Die Recherche-Oberfläche wird als Applikation in die Software Mapbender integriert, welche ebenfalls als Open-Source-Software zur Verfügung steht. Der Einsatz von Mapbender ist dadurch begründet, das parallel zur Entwicklung des Geodatenkatalog-DE der Aufbau eines Mapservers stattfindet. Hier dient Mapbender als Diensterepository und Konfigurationsoberfläche der Dienste, die über den Mapserver bereitgestellt werden. Diese Dienste sollen ebenfalls in die Suche des Geodatenkatalog-DE integriert werden, sodass eine Rechercheoberfläche in Mapbender sich anbietet.

---

[5] http://www.opengeospatial.org

Als Datenbankkomponente wird die Open-Source-Software PostgreSQL[6] eingesetzt, da diese relationale Datenbank von beiden Softwarekomponenten des Geodatenkatalog-DE unterstützt wird und somit nur ein Datenbankinstanz betrieben werden muss. Für die Software Mapbender ist die PostgreSQL-Erweiterung PostGIS notwendig.

Weiterhin wird im Geodatenkatalog die Mapbender-Applikation MeTaDor zur Erfassung und Bearbeitung von Metadatensätzen eingesetzt. Mit dieser Anwendung erfasste Metadaten lassen sich über ein Harvesting-Mechanismus direkt in den Geodatenkatalog-DE überführen.

4 Umsetzung

4.1 Harvesting

Das Harvesting der angeschlossenen Metadatenkataloge erfolgt im Geodatenkatalog-DE über die Software GeoNetwork. Diese kann über diverse Katalogschnittstellen externe Metadatenkataloge harvesten, wobei für den Geodatendatenkatalog-DE die CSW-Schnittstelle die präferierte Lösung ist. Unterstützt wird der CSW in den Versionen 2.0.1 und 2.0.2. Die einzubindenden Metadatenkataloge der GDI-Partner werden standardkonform (ISO, OGC) über den Capabilities-Request des CSWs angefragt und das Capabilities-Dokument, welches die Eigenschaften des Services beinhaltet, wird von der Software ausgelesen. Über die GetRecord bzw. GetRecordById-Operation werden die einzelnen Metadatensätze in den Geodatenkatalog-DE übernommen. (siehe Abb.1) In GeoNetwork kann auf die CSW Operationen über POST, GET methods und SOAP encoding zugegriffen werden.

Über die UUID des Metadatensatzes wird überprüft, ob dieser bereits im System vorliegt. Falls dies der Fall ist, wird der Metadatensatz nur dann erneut geladen, wenn sich das Datum der letzten Änderung geändert hat. Dadurch wird verhindert, dass Metadatensätze mehrmals im Geodatenkatalog-DE vorhanden sind. Unveränderte Metadatensätze werden nicht erneut geladen, dadurch wird der Traffic auf dem Server erheblich reduziert.

Die Harvesting-Aufgaben können in GeoNetwork konfiguriert und administriert werden. Über das Harvesting-Management kann der GeoNetwork-Administrator Kataloge hinzugefügten, bearbeiten oder entfernen. Durch das Entfernen eines Kataloges werden auch alle Metadatensätze dieses Kataloges aus dem Geodatenkatalog-DE entfernt. Das Hinzufügen eines Kataloges erfolgt durch die Angabe des Capabilities-Request und der Benennung des Dienstes. Weiterhin können für nicht frei zugängliche Dienste die Zugangsdaten eingetragen werden. Sogenannte Suchkriterien schränken das Harvesting Ergebnis ein. Diese Filter können für die Suchelemente “Free Text”, “Title”, “Abstract” und “Subject” gesetzt werden. Zudem kann ein zeitlicher Intervall angegeben werden, der festlegt, wie oft (bzw. wie regelmäßig) der Katalog zum Harvesting angefragt werden soll. Bestimmte Kataloge können auch einzelnen Gruppen zugeordnet werden, sodass nur Mitglieder dieser Gruppe Zugriff auf die Metadatensätze haben. Diese Einträge bzw. Einstellungen werden gespeichert und der Katalog ist in der Liste der eingebundenen Kataloge im Harvesting-

[6] http://www.postgresql.org

Management sichtbar. Durch das Aktivieren des Kataloges in dieser Liste erfolgt das Harvesting der Metadatensätze dieses Kataloges.

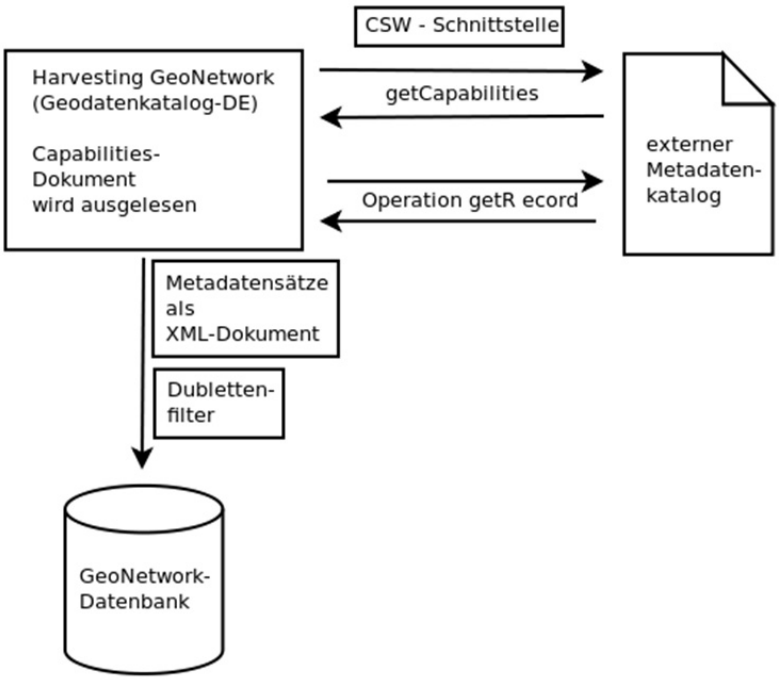

Abb. 1: Harvesting im Geodatenkatalog-DE

Probleme beim Harvesting entstehen der Erfahrung nach da, wo Kataloge die OGC-konformen Anfragen von GeoNetwork nicht klar beantworten kann. Anscheinend wird der CSW-Standard von einigen Metadatenkatalogen unterschiedlich interpretiert. Dies betrifft allerdings nur sehr wenige Metadatenkataloge.

4.2 Recherche

Die Rechercheoberfläche bietet dem Nutzer die Möglichkeit, gezielt nach Geoinformationen zu suchen. Hierzu werden über die CSW-Schnittstelle die Suchmechanismen von GeoNetwork genutzt. Geonetwork nutzt einen integrierten Lucene-Index zur Suche, welcher nach jedem Harvesting neu aufgebaut und optimiert wird. (über die Administrationsoberfläche von GeoNetwork kann dieser auch jederzeit manuell neu aufgebaut und optimiert werden) Dieser gewichtet die Ergebnisse und gibt diese an die Rechercheoberfläche mit einem entsprechenden Ranking ab. Über die Suche werden entsprechend der Suchparameter diverse Felder des Lucene-Index abgefragt. Beispielsweise erfolgt eine einfache Suche (Einfeldsuche) über das Feld „any“ des Indexes. Dieses Feld enthält u.a die Begriffe aus den Titelangaben und Kurzbeschreibungen.

In der Rechercheoberfläche werden die Ergebnisse (Treffer) als sortierbare Liste dargestellt. Die Ansicht der einzelnen Ergebnisse enthält definierte Informationen, welche aus den x-pathes des XML-Dokumentes des einzelnen Metadatensatzes ausgelesen werden. Diese XML-Dokumente werden in der Datenbank von GeoNetwork in der Tabelle metadata gespeichert. Über einen Link gelangt der Nutzer zu einer ausführlicheren Darstellung des Metadatensatzes. Auch diese Informationen werden aus dem XML-Dokument extrahiert.

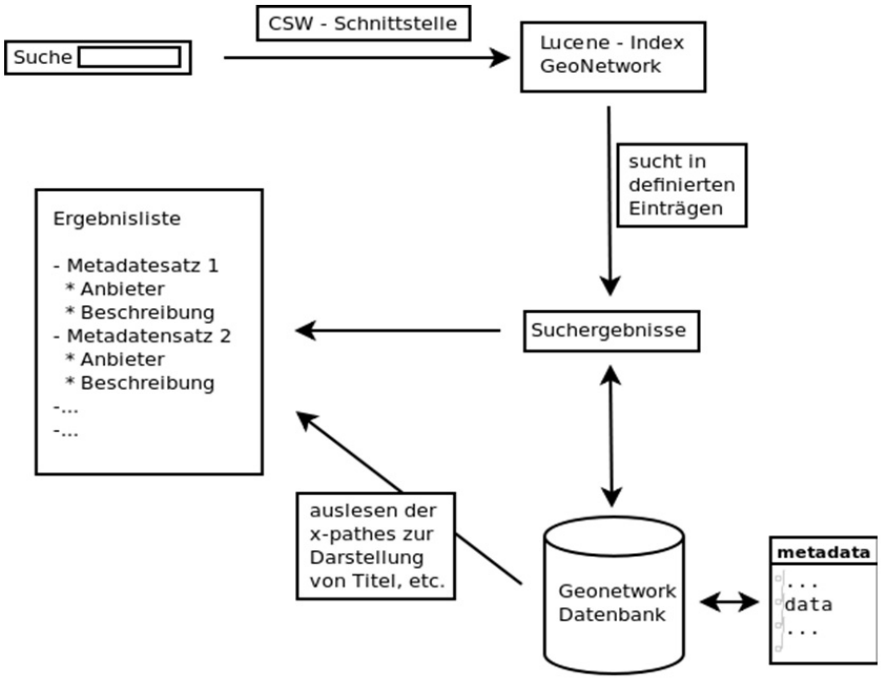

Abb. 2: Recherche im Geodatenkatalog-DE

Zusätzlich wird eine Suche nach Geodiensten, die in Mapbender registriert sind, angestoßen. Diese Suche erfolgt in der Mapbender-Datenbank und durchsucht die Tabelle der Metadaten der registrierten Dienste. Bei der Registrierung eines Dienstes wird das Capabilities-Dokument ausgelesen und die Metadateninformationen in die Datenbank eingetragen. Das Ergebnis wird ebenfalls als Liste zurückgegeben.

5 Anpassungen

Im Zuge des Aufbaus des Geodatenkatalog-DE wurden einige Softwareanpassungen vorgenommen. Diese Weiterentwicklungen fließen, wie bei Open-Source-Projekten üblich, wieder dem Projekt zu und sind somit in nachfolgenden Versionen integriert. Zwei Anpassungen in GeoNetwork seien hier erwähnt, da diese das Harvesting betreffen.

- Anpassung des Schedulers: War es bisher nur möglich, das Intervall der einzelnen Harvesting-Anfragen anzugeben (in Tagen/Minuten/Stunden) können nun konkrete Termine, Uhrzeiten und Intervalle angegeben werden (z. B. „alle zwei Tage um 14.30 Uhr“).
- Erweiterung der Logging-Ausgabe für Harvesting-Jobs über die Oberfläche: Über die Oberfläche ist es nun für den GeoNetwork-Administrator möglich, ausführlichere Informationen zum Harvesting eines Kataloges abzurufen. Zudem werden die Informationen historisiert. Besonders, wenn ein Harvesting erfolglos ist, können diese Informationen zur ersten Fehleranalyse genutzt werden.

6 Fazit und Ausblick

Mit dem Geodatenkatalog-DE ist ein wichtiger Bausteine der GDI-DE entwickelt worden, der einen zeitgemäßen und intuitiven Zugang zu den Meta- und Geodatenbeständen der GDI-DE bieten. Besonders als zentrale Recherchemöglichkeit der GDI-DE, welche durch das Harvesting über die standardkonforme Schnittstelle CSW realisiert wird, bietet der Geodatenkatalog-DE dem Nutzer den Vorteil über eine Anwendung in sehr vielen dezentralen Datenbeständen nach Geoinformationen zu suchen. Auch in Zukunft ist so ein effizientes Suchen und Nutzen von Geodaten immer besser möglich, da der Katalog durch die GDI-DE und seinen Partnern aus Verwaltung und Wirtschaft weiter ausgebaut wird.

Der Einsatz von freier Software ermöglicht die einfache Anpassung an die Anforderungen und Wünsche der Koordinierungsstelle und bietet größtmögliche Flexibilität bei der Einbindung der neuen Dienste in eine Vielzahl von Anwendungen und Portale.

Weiterführende Hinweise

- http://lucene.apache.org/java/docs/index.html.
- http://www.geonetwork-opensource.org/manuals/2.6.3/users/index.html.
- http://www.geonetwork-opensource.org/manuals/2.6.3/developer/index.html.
- http://www.mapbender.org.
- RICHTLINIE 2007/2/EG DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 14. März 2007 zur Schaffung einer Geodateninfrastruktur in der Europäischen Gemeinschaft (INSPIRE).