
PortalU[®] und InGrid[®] – Werkzeuge zur Erstellung, Recherche und Verteilung von Metadaten

Martin KLENKE, Thomas VÖGELE, Fred KRUSE, Hanno LEHMANN

Zusammenfassung

Metadaten sind für ein sinnvolles und nachhaltiges Management von (Geo-)Daten unerlässlich. Datenkataloge spielen daher beim Aufbau von Geodateninfrastrukturen, wie der zukünftigen europäischen INSPIRE-Infrastruktur (Infrastructure for Spatial Information in Europe), eine zentrale Rolle. Im aktuellen Umfeld von Metadatenerfassung und –recherche existieren zwei wesentliche Problemfelder: Zum Einen ist ein großer Teil vorhandener analoger und digitaler Umweltdaten im Bereich der öffentlichen Verwaltung nicht in ausreichender Weise standard-konform mit Metadaten dokumentiert. Wenngleich sich das Erfordernis angemessener Metadatendokumentation auf konzeptioneller Ebene langsam durchsetzt, zeigt die Praxis, dass die Umsetzung vor erheblichen Problemen steht, insbesondere der Tatsache, dass Metadatenerfassung zeitlich und inhaltlich aufwendig ist und dem Datenhersteller, der seine Daten „kennt“, keinen unmittelbaren Nutzen verspricht. Der Entwicklung von Werkzeugen, die Datenbereitsteller bei der Erfassung bestmöglich unterstützen und deren Erfassungsaufwand reduzieren, kommt daher besondere Bedeutung zu.

Ein zweites Problemfeld existiert im Bereich Interoperabilität. Vernetzte Geodateninfrastrukturen erfordern Schnittstellen, die eine katalogübergreifende Recherche ermöglichen. Auf konzeptioneller Ebene stehen für den Austausch von ISO 19115 und ISO 19119 konformen Metadaten die Catalog Service Web-Spezifikationen (CSW) und Empfehlungen des Open Geospatial Consortium (OGC 2007, 2005) zur Verfügung. Auf diese wird in zahlreichen aktuellen Konzeptionen, z. B. im Umfeld von GDI-DE (Geodateninfrastruktur Deutschland) und INSPIRE, Bezug genommen. Aktuelle Testbeds zeigen allerdings, dass noch erhebliche Probleme mit der Kommunikation unterschiedlicher CSW-Implementierungen existieren. Schwierigkeiten bei der Umsetzung mehrstufiger Kaskaden und mit der Performanz des Protokolls im Allgemeinen wurden noch nicht ausreichend diskutiert.

In unserem Beitrag möchten wir, mit Bezug auf den ersten der oben skizzierten Problembereiche, die derzeit in Entwicklung befindliche Weberfassungs-Komponente des Umweltportals Deutschland PortalU[®] vorstellen. Sie wird ab 2008 den bislang in der deutschen Umweltverwaltung zur Metadatenerfassung eingesetzten Windows-UDK 5.0 (SWOBODA et al. 1999) ersetzen. Als Diskussionsbeitrag zum Thema Interoperabilität stellen wir im zweiten Teil verschiedene Schnittstellen und Kommunikationskonzepte von InGrid[®] vor. InGrid[®], die Software hinter PortalU[®], verfolgt einen streng dezentralisierten Ansatz der Abfrage verteilter Datenquellen und Indizes auf Basis von P2P-Kommunikationstechniken.

1 Umweltdatenkatalog und PortalU[®]

In Deutschland spielt der Umweltdatenkatalog (UDK) eine zentrale Rolle bei der Referenzierung und Beschreibung von Daten, Karten und Dokumenten im Umweltbereich (SWOBODA 2000). Seit mehr als 10 Jahren ist der UDK bei den Umweltbehörden des Bundes und der Länder im Einsatz und hat sich in diesem Umfeld zu einem Quasi-Standard zur Beschreibung von Umweltinformationen entwickelt. Das Datenmodell des UDK berücksichtigt die internationalen Standards ISO 19115 und ISO 19119 (KARSCHNICK 2003).

Als technisches System wurde der UDK bisher mit einer Online Komponente für Recherche und Visualisierung (WWW-UDK), sowie einer Desktop-Applikation (Windows-UDK) für die Eingabe von Metadaten und für das Katalogmanagement betrieben. Seit Mai 2006 ist die Online-Recherche-Komponente des UDK in PortalU[®] integriert. Seither können unter dem gemeinsamen Dach von PortalU[®] nicht nur Webseiten und Fachinformationssysteme von Behörden, sondern auch die UDKs des Bundes und der deutschen Länder durchsucht werden.

1.1 Eingabe und Pflege von Metadaten

Um die Dateneingabe und den kompletten Workflow, von der Metadatenerfassung und Pflege über die Qualitätssicherung (QS) bis hin zur Datenrecherche, über das Internet abwickeln zu können, wird derzeit eine neue Softwarekomponente entwickelt: InGrid[®] 1.1. Diese wird ab 2008 die bisherige Desktop-Erfassungssaplikation ersetzen.

Die Entscheidung, zukünftig ein internet-basiertes Tool für Erfassung, Pflege und QS zu verwenden, basiert auf der langjährigen Erfahrung mit einem Katalogmanagement, in dessen Rahmen verteilt gepflegte Kataloge mit Hilfe eines definierten Aktualisierungszyklus¹ in einen zentralen Katalog überführt wurden und dieser dann wieder in die nachgeordneten Installationen verteilt wurde. Dieses Vorgehen hat sich in der Praxis sowohl technisch als auch organisatorisch als wenig robust erwiesen. Das neue Werkzeug wird daher mit Ajax²-Technologie als benutzerfreundliche und browserbasierte Internetapplikation umgesetzt, die unabhängig vom Standort des Metadatenautors das Schreiben in einen gemeinsamen Katalog ermöglicht. Bei der Entscheidung für Ajax, das aktivierte Javascript im Browser bedingt, setzten sich Gesichtspunkte der Ergonomie gegen die der Barrierefreiheit (im Sinne der BITV) durch: Die Metadatenautoren erwarten nach langjähriger Erfahrung mit einer komfortablen, performanten Desktopanwendung ein ähnlich flüssiges Arbeiten auch unter einer Webapplikation. Da die Erfassung/Qualitätssicherung lediglich einen autorisierten Nutzerkreis betrifft, wurden einige BITV-Belange zugunsten der Usability zurück gestellt.

1.2 InGrid 1.1[®] Metadatenerfassungskomponente

Im Folgenden werden einige innovative Aspekte der Anwendung vorgestellt.

¹ <http://www.portalu.de>

² [http://de.wikipedia.org/wiki/Ajax_\(Programmierung\)](http://de.wikipedia.org/wiki/Ajax_(Programmierung)) – 26.05.2007

Das UDK-Datenmodell berücksichtigt, wie oben angesprochen, die ISO 19115 und 19119 und geht in Teilen darüber hinaus. Insgesamt existieren mehr als 450 Eingabefelder. In der Praxis ist allerdings meist nur ein geringer Teil der vorhandenen Metadatenelemente erforderlich, um ein Datenobjekt im Sinne der Standards ausreichend zu beschreiben. Die Mindestanforderungen der ISO-Standards, der so genannte ISO-Core, umfassen daher weniger als 20 Metadatenelemente. Um die Usability zu unterstützen und unnötige Komplexität zu vermeiden besteht in InGrid® 1.1 grundsätzlich die Möglichkeit der Wahl zwischen einer vollständigen und einer auf die Core-Elemente reduzierten Darstellung. Zusätzlich kann der Katalogadministrator festlegen, welche Elemente über die Eingabeoberfläche per Voreinstellung sichtbar sind. Das „Ausfalten“ der Ansichten ist bei Bedarf für einzelne Elementbereiche oder global möglich. Damit bietet sich die Möglichkeit, die Metadatenerfassungsoberfläche katalogweise an die Bedürfnisse spezifischer Landes-Metadatenprofile anzupassen.

Schlagworte werden zur inhaltlichen Beschreibung und Kategorisierung von Datensätzen genutzt und sind für den Erfolg einer Datenrecherche mitentscheidend, insbesondere wenn über vernetzte Kataloginfrastrukturen gesucht wird. Um das Finden geeigneter Schlagworte zu unterstützen wird ein Verschlagwortungsassistent eingesetzt, der über einen externen Webservice, den Semantic Network Service³ (SNS) des Umweltbundesamts, Metadatensätze semantisch analysiert und auf dieser Grundlage Schlüsselwörter vorschlägt. Die zugrunde liegenden Fach- und Geothesauri unterstützen die inhaltliche Qualität der vergebenen Schlagworte und fördern die semantische Homogenität des Gesamtdatenbestands.

Die Qualitätssicherung neu eingepflegter sowie die Kontrolle bereits vorhandener Metadaten, ist erforderlich um sicherzustellen, dass ein Datenkatalog in sich konsistent und homogen ist und bleibt. InGrid® 1.1 sieht hierfür eine zweistufige Qualitätskontrolle vor: Neu eingepflegte Metadaten werden, bevor sie im Datenkatalog abgespeichert werden können, einer manuellen Qualitätskontrolle unterzogen. Dabei werden die Metadaten automatisch an eine mit der Qualitätssicherung beauftragte Person weitergeleitet. Erst nach Durchlaufen der Qualitätssicherung können die Daten endgültig abgespeichert und der externen Recherche zur Verfügung gestellt werden. Als weitere QS-Maßnahme werden, nach einer frei zu definierenden Verfallszeitspanne, alle Metadaten automatisch dem jeweiligen Datenverantwortlichen zur erneuten Kontrolle und Aktualitätsprüfung vorgelegt. Interne Mechanismen stellen sicher, dass die ISO-Pflichtfelder bereits vor der Weiterleitung eines Objekts an die QS befüllt wurden. Der QS-Zyklus ist für einen Katalog optional und nicht verpflichtend.

Zahlreiche Institutionen der öffentlichen Verwaltung nutzen Komponenten aus ESRI's Produktfamilie um Geodaten zu prozessieren und zu dokumentieren. Daher wird es für diesen Anwenderkreis eine XML-Schnittstelle geben, über die Metadaten, die mit dem ESRI-Metadateneditor gepflegt werden, direkt in einen UDK überführt werden können. Die Befüllung der ISO-Pflichtfelder in den zu importierenden Datensätzen wird dabei über einen internen QS-Mechanismus sichergestellt.

³ <http://www.semantic-network.de>

2 Interoperabilität und Schnittstellen

2.1 Architektur

Die schraffiert hinterlegten Komponenten in Abbildung 1 geben einen Überblick über die InGrid®-Softwarearchitektur. Als zentraler Informationsbroker fungiert der Informationsbus („iBus“), an den über „iPlugs“ externe Datenquellen, z. B. Fachinformationssysteme, Datenbanken und Suchmaschinenindizes angeschlossen werden können.

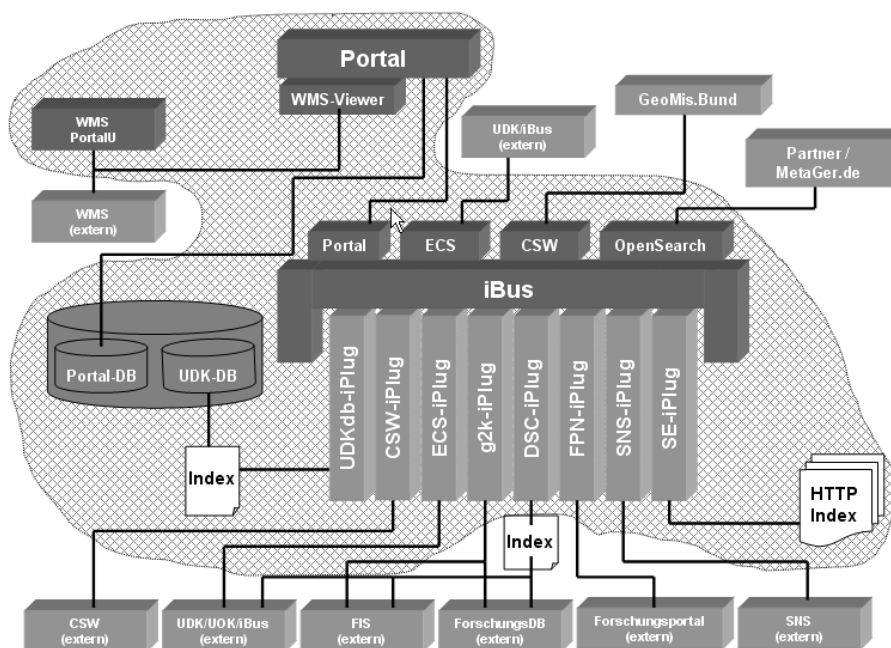


Abb. 1: InGrid®-Architektur

Der iBus vermittelt zwischen den angeschlossenen Datenquellen und den Systemen, die auf diese zugreifen möchten: Intern das Portal, über das ein Nutzer Suchanfragen stellen und, entsprechende Berechtigungen vorausgesetzt, administrative Aufgaben wahrnehmen kann, sowie externe Systeme, die PortalU® abfragen, in der Abbildung z. B. GeoMIS.Bund über die CSW-Schnittstelle und die Metasuchmaschine MetaGer.de über die OpenSearch-Schnittstelle⁴.

⁴ <http://www.metager.de>, <http://www.opensearch.org> – 26.05.2007

2.2 Komponenten und Kommunikation

Sämtliche PortalU®-Komponenten sind auf Basis von Java implementiert und basieren auf Open-Source-Software (OSS). Tabelle 1 fasst die wesentlichen Komponenten zusammen.

Tab. 1: InGrid® OSS-Komponenten und Projektadressen

| | | |
|---------------|---------------|---|
| Suchmaschine, | Nutch | http://lucene.apache.org/nutch |
| Indexierung | Apache Lucene | http://lucene.apache.org |
| Portal | Jetspeed | http://portals.apache.org/jetspeed-2 |
| WMS | UMN Mapserver | http://mapserver.gis.umn.edu |
| WMS Viewer | Mapbender | http://www.mapbender.org |
| Datenbank | MySQL | http://www.mysql.com |
| Kommunikation | JXTA | http://www.jxta.org |

Formatiert: Englisch
(Großbritannien)

Die jxta-P2P-Plattform ermöglicht die physikalische Verteilung von iPlugs. Zum Anschluss einer Datenquelle kann der verbindende Plug sowohl lokal am iBus als auch extern an der Datenquelle eingerichtet werden. Die Kommunikation kann firewall-freundlich über http oder, wenn möglich, über das performantere tcp-Protokoll erfolgen.

Ein zentrales Problem von verteilten Dateninfrastrukturen sind mehrstufige Kaskaden von An-/Abfragesystemen, die Dubletten erzeugen können und häufig wenig performant arbeiten. Letzteres fällt besonders stark ins Gewicht, wenn die Kommunikation zusätzlich über wenig performante Protokolle, wie z. B. XML/SOAP, erfolgt. Dies kann zu Antwortzeiten im Minutenbereich führen, die Nutzer in Zeiten von Google und anderen performanten, index-basierten Suchmaschinen kaum akzeptieren. Ein Beispiel ist der CSW-basierte Geodatenkatalog von GeoPortal.Bund⁵, der, obwohl nicht mehrfach kaskadiert, einen Timeout von zwei Minuten setzt. Die InGrid®-Architektur erlaubt den Verzicht auf mehrstufige Hierarchien, da jede Datenquelle mit ihrem iPlug direkt an beliebig viele iBusse angeschlossen werden kann. Die P2P-Technologie automatisiert dabei die Kontaktaufnahme sowie die Kommunikation zwischen iBus und iPlugs weitgehend.

2.3 Schnittstellen und Interoperabilität

PortalU® verfügt mit dem „Data Source Client“ (DSC) über ein generisches Schnittstellenmodul das es ermöglicht, Datenbanken und Fachinformationssysteme komfortabel und ohne Eigenentwicklung auf Anbieterseite an das Portal anzubinden. Datenbankinhalte können dabei sowohl in der PortalU®-Detailansicht dargestellt, als auch aus der PortalU®-Ergebnisliste direkt per URL aus einem Fremdsystem geladen werden, sofern das angeschlossene System über eine eigene Ergebnisanzeige verfügt. Damit können auch Teile des sogenannten „Hidden Web“ einfach und kostengünstig erschlossen werden. Der DSC erstellt zudem einen Index der gemappten Datenbankinhalte, so dass diese in der PortalU®-Trefferliste suchanfragen-spezifisch und datenquellen-übergreifend nach Relevanz sortiert werden können. Zum Suchanfragezeitpunkt wird, um die Ergebnisliste aufzubauen,

⁵ <http://geoportal.bkg.bund.de> – 26.05.2007

lediglich auf den Index zugegriffen, was kurze Antwortzeiten ermöglicht. Erst bei der Anforderung eines Detailergebnisses wird auf die Quelldatenbank durchgegriffen.

Neben dem DSC verfügt InGrid[®] über verschiedene XML-basierte An- und Abfrageschnittstellen: Über die CSW-Schnittstelle wird PortalU[®] einerseits die Inhalte der angeschlossenen UDKs für GeoPortal.Bund (und langfristig INSPIRE) verfügbar machen. Andererseits sollen darüber externe Geodaten-Kataloge an PortalU[®] angebunden werden. Die ECS-Schnittstelle (Environmental Catalog Service) ist eine Weiterentwicklung der UDK-SOAP-Schnittstelle und bildet das gesamte fachliche UDK-Datenmodell ab. Sie soll sich zukünftig zu einer Standardschnittstelle für den Austausch von Umwelt-Metadaten entwickeln.

3 Zusammenfassung

In der nächsten Ausbaustufe wird PortalU[®] ab 2008 mit InGrid 1.1[®] über eine leistungsfähige Web-Erfassungskomponente verfügen, die sich nahtlos in die bestehende InGrid[®]-Architektur einfügt. Die Software steht über eine Verwaltungsvereinbarung dem Bund und den deutschen Ländern zur Verfügung und stellt die deutsche Umweltverwaltung, durch innovative Architektur- und Schnittstellenkonzepte, im Hinblick auf die Anforderungen der Umweltinformationsgesetze und die geplanten nationalen und internationalen Dateninfrastrukturen, zukunftssicher auf.

4 Literatur

- KARSCHNICK O. et al. (2003), The UDK and ISO 19115 Standard. - In: Proceedings of the 17th International Conference "Informatics for Environmental Protection". Cottbus, Germany.
- OPEN GEOSPATIAL CONSORTIUM, INC (2007), OpenGIS[®] Catalogue Service Implementation Specification Version 2.0.2. - Web: http://portal.opengeospatial.org/files/?artifact_id=20555 (25.04.2007).
- OPEN GEOSPATIAL CONSORTIUM, INC (2005), ISO19115/ISO19119 Application Profile for CSW 2.0. Version 0.9.2 - Web: http://portal.opengeospatial.org/files/?artifact_id=6495 (25.04.2007).
- SWOBODA, W. et al. (2000), Harmonisierter Zugang zu Umweltinformationen für Öffentlichkeit, Politik und Planung: Der Umweltdatenkatalog UDK im Einsatz. - In: K. Tochtermann, W.-F. Riekert (Eds). Proceedings of the 14th International Symposium „Computer Science for environmental Protection“, Bonn, Germany.
- SWOBODA W. et al. (1999), The UDK Approach: the 4th Generation of an Environmental Data Catalogue Introduced in Austria and Germany. - Proceedings of the 3rd IEEE Meta-Data Conference, Bethesda, Maryland, USA.