Neue Wege für Metadaten

Arnulf Christl, Metaspatial

Zusammenfassung

Im Vortrag werden zunächst die Grundlagen der Metadatenverarbeitung vorgestellt. Mit Anleihen aus dem Kontext der Linguistik werden Syntax und Semantik von Metadaten in der räumlichen Datenverarbeitung erläutert. Es folgt eine kurze Übersicht zur Bedeutung von Ontologien und es wird auf die Pragmatik als dritte Disziplin der Semiotik verwiesen. Aus dem Mangel an Pragmatik können die aktuellen Schwächen von Metadaten-Formaten und Katalogen abgeleitet werden. Im Ausblick wird erläutert, wie der grundlegenden Mangel an semiotischer Pragmatik überwunden werden kann. Einfache Beispielen sollen helfen, den linguistischen Fachjargon in einen räumlichen Kontext zu setzen.

Einführung

Aktuelle Richtlinien der Geodatenverarbeitung (z.B. INSPIRE[1]) schreiben vor, Metadaten zu erfassen, zu kategorisieren und über das Internet bereitzustellen. Hierfür werden typischerweise Katalogdienste eingesetzt. In der Theorie können Anwender über Anwendungen, die diese Katalogdienste verwenden, Kartendienste und Geodaten finden, um sie für eigene Zwecke zu nutzen. Deshalb werden derzeit mit viel Aufwand und durch gesetzliche Richtlinien wie INSPIRE teilweise verpflichtend, umfangreiche Metadatenbestände aufgebaut.

In der Praxis zeigt sich jedoch, dass diese Metadaten aus verschiedenen Gründen nicht besonders gut geeignet sind, um Geodaten und Dienste zugänglich zu machen. Die in den Metadaten hinterlegten Informationen sind zu spärlich und die Kataloge zu statisch und wenig flexibel, um den Anforderungen der Anwender zu genügen. Ein anderes Problem liegt in den unterschiedlichen Herangehensweisen und der Motivation von Datennutzern und Anwendern, ein weiteres in mangelndem Verständnis der verwendeten Technologien.

Die OSGeo[2] erkundet neue Wege, um diese Kluft mit web-basierten Technologien und neuen Kommunikations-Möglichkeiten zu überbrücken. Dabei geht es nicht nur um Software, und Resourcen-basierte (REST[3]) Architektur-Modelle, sondern vor allem um Online Kommunikationsmedien. Diese bilden gemeinsam neue Kulturwerzeuge auf Basis des Internet, deren Erforschung noch in den Kinderschuhen steckt.

---

1 Infrastructure for Spatial Information in the European Community http://www.inspire-geoportal.eu/

2 Open Source Geospatial Foundation, http://wiki.osgeo.org/wiki/INSPIRE

3 Representational State Transfer http://wiki.osgeo.org/wiki/REST

Metadaten

Metadaten enthalten beschreibende Informationen über Daten. In der räumlichen Datenverarbeitung ist die Verfügbarkeit von Metadaten eine grundsätzliche Voraussetzung für die sinnvolle Anwendung und Nutzung von Geodaten, Diensten und Anwendungen.

Zwei Beispiele für unterschiedliche Metadaten

Metadaten zu einem Orthophoto können unter anderem folgende Informationen bereitstellen (die Liste ist unvollständig):

- Räumlicher Ausschnitt
- Koordinatensystem, Projektion
- Format oder Zugriffsmöglichkeiten
- Datum der Aufnahme
- Auflösung des Originalbildes
- Farbkanäle
- Aufnahmegerät
  - Digital
  - Analog
- Bearbeitungsschritte
  - Ausschnitt
  - Entzerrung
  - Schattenaufhellung
  - Kontrast- und Farbanpassung
  - etc.

Andere Geodaten benötigen andere Informationen, für die Karten eines Stau-Informationsdienstes wären unter anderem folgende Aspekte wichtig:

- Koordinatensystem, Projektion
- Datum der letzten Aktualisierung
- Format oder Zugriffsmöglichkeiten
- Ursprung der geometrischen Grundlage
- Aktualität der geometrischen Grundlage
- Aufnahmeart
  - Datenerhebung durch Verkehrsüberwachung (amtlich)
  - Meldungen durch Autofahrer (freiwillig, verifiziert)
- Prognosestatus
  - Statistische Auswertung
  - Berücksichtigung von Baustellen
  - etc.

Die meisten dieser Metadaten liegen heute bereits digital vor. Statt sie in einer vorgegebenen Syntax hierarchisch strukturiert für einen Metadatenkatalog "abzutippen", sollten sie einfach offen gelegt werden. Das heißt nicht, dass alle Daten sofort gemeinfrei werden, sondern le-

diglich, dass dem Suchenden die Möglichkeit an die Hand gegeben wird selbst herauszufinden, welche Metadaten er braucht.

Syntax und Semantik von Metadaten

Die Struktur von Metadaten können unter Aspekten der Syntax und der Semantik analysiert werden. Die Syntax beschreibt Geodaten rein **formal**, während sich die Semantik auf deren **Bedeutung** bezieht. Beide sind wichtige Voraussetzungen für die Nutzung von Geodaten und Diensten im Netz. Syntax und Semantik sind Disziplinen der Semiotik, die durch die Lehre von den Beziehungen der Zeichen zu den Zeichenbenutzern (Pragmatik [1]) vollendet wird; und genau hier liegt das Problem.

Doch zunächst zur Syntax: Ein Teil der oben beschriebenen Metadaten kann so harmonisiert werden, dass sie für beide Beispiele im gleichen Format angegeben werden können. Diese Informationen beziehen sich dann auf die Syntax. Dazu zählen das Koordinatensystem und die Projektion, die aus einer definierten Liste ausgewählt, oder direkt als Funktionen mit Parametern definiert werden können. Das Format oder die Zugriffsmöglichkeiten können ebenfalls in einer für beide Geodaten gleichen Form beschrieben werden. Hier können Verweise auf klare Definitionen wie z.B. Standards gegeben werden, seien es de-jure Standards der ISO[4], offene Industriestandards des OGC[5] oder proprietäre Formate wie sie von Bing, Yahoo oder Google vorgegeben werden, und für die technische Beschreibungen vorliegen. Diese Informationen beziehen sich auf die Syntax der Geodaten. Die Syntax beschreibt sozusagen die Grammatik. Ein Beispiel aus der Linguistik verdeutlicht diesen Aspekt: Der Satz "Ich bin ein Tisch, deshalb werden Schwäne grün" ist rein syntaktisch richtig, auch wenn ich in Wirklichkeit kein Tisch bin, noch dies grüne Schwäne bedingen würde.

Semantische Informationen sind ungleich komplexer und weniger greifbar, da sie weniger gut strukturiert werden als die Syntax. Semantische Informationen beschreiben die Daten auf inhaltlicher Ebene und beziehen sich auf Sinnzusammenhänge. Der Sinnzusammenhang ist ein wichtiges Kriterium zur Auffindbarkeit von Geodaten und Diensten. Die mangelnde Abbildung von Sinnzusammenhängen in technischen (syntaktischen) Metadatenbeschreibungen führt dazu, dass Katalogsuchen oft nicht zu den gewünschten Ergebnissen führen.

Ontologien

Die Ontologie ist die Lehre vom "Sein der Dinge". Eine Ontologie beschreibt die Beziehungen von Sinnzusammenhängen und bringt sie in einen definierten Kontext. Erst das Akzeptieren und die Anwendung einer gemeinsamen Ontologie oder eines Sinnzusammenhangs ermöglicht Kommunikation. Gruninger und Lee [2] unterscheiden drei Anwendungsfelder: *Kommunikation*, *automatisches Schließen* und *Repräsentation sowie Wiederverwendung von Wissen*. Sollen zwei Programme (z.B. Web-Suchmaschinen oder Software-Agenten) miteinander kommunizieren, so müssen sie entweder selbst die Interpretationsvorschrift für die Daten in

4 International Organization for Standardization, http://www.iso.org/

5 Open Geospatial Consortium, http://www.opengeospatial.org/

sich tragen (sind also datenabhängig), oder aber sie liefern diese in Form von Metadaten einer beiden Seiten zugänglichen Ontologie mit. Hier werden Metadaten als Form der Kommunikation verstanden, mit Hilfe derer Metadaten überhaupt erst interpretierbar werden.

Ontologie werden ebenfalls benötigt, um die Regeln einer Syntax zu kommunizieren. Bezogen auf die oben gegebenen Beispiele beschreibt die Ontologie des OGC verschiedene Dienste (Services) in mehreren "Lehren" (Standard-Dokumenten). Dabei werden gemeinsame Sinn-Grundlagen in übergeordneten "Lehren" oder Standards zusammengefasst. Um beim Beispiel zu bleiben, schreibt der "OGC Commons Standard" vor, dass alle drei großen Dienste (WMS[6], WFS[7], WCS[8]) ein Capabilities-Dokument bereitstellen müssen. Der Commons-Standard verweist zusätzlich auf die EPSG (European Petroleum Survey Group), in der die Beschreibung von Koordinatensystemen und Projektionen definiert sind. In diesen wird angenommen, dass klar ist, wie eine Rechenoperation ausgeführt wird oder was eine Gleitkommazahl ist, hier wird implizit Bezug auf Grundrechenoperationen genommen.

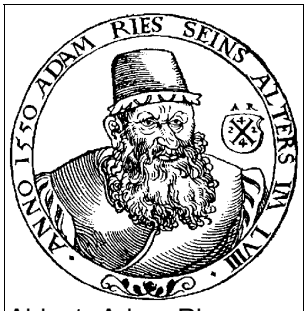

Abb. 1: Adam Rise

Wie ist nun die Ontologie der Grundrechenoperationen zustande gekommen und warum addieren wir heute mit arabischen Zahlen? Weil *"2 + 2* ***nach Adam Rise*** *4 macht"* [3]. Adam Rise (siehe Abbildung 1) hat mit seinen grundlegenden Werken in deutscher Sprache (damals war Latein die Sprache der Gelehrten) entscheidend dazu beigetragen, dass die römischen Zahlzeichen durch die nach dem Stellenwertsystem strukturierten indisch-arabischen Zahlzeichen ersetzt wurden. Er hat diese Ontologie bekannt gemacht und damit Akzeptanz für die Anwendung geschaffen. Wir können festhalten, dass sowohl Syntax, als auch Semantik auf Ontologien basieren.

Ontologien, auf der die Semantik eines Geodatensatzes oder Dienstes aufbauen können, sind oft fachspezifisch. Ein Beispiel ist die ozeanographische Forschung, deren Ontologien von mehreren eigens geschaffenen Organisationen [4], [5] aufgebaut und mit modernen Kommunikationsmethoden (siehe Abbildung 2) diskutiert, gepflegt und im besten Fall auch genutzt werden.

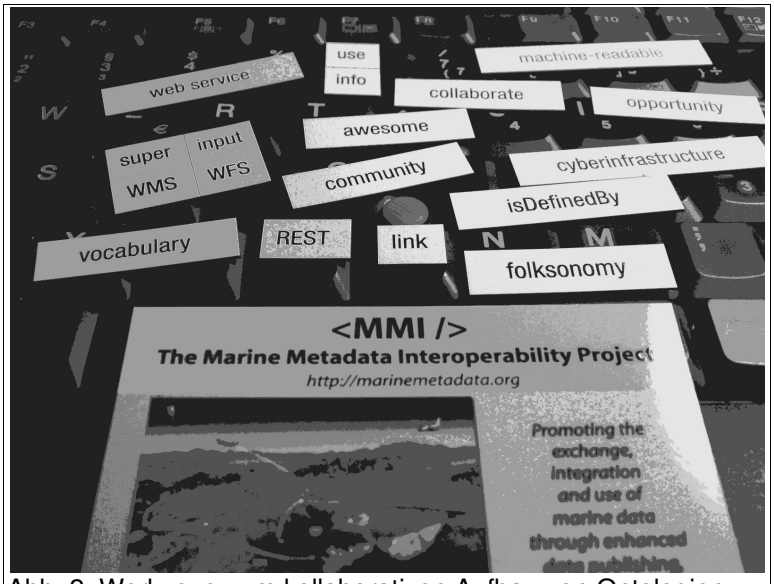

Abb. 2: Werkzeug zum kollaborativen Aufbau von Ontologien

Die Ontologien der Ozeanographie verweisen auf andere,

6 OGC Web Map Service, Karten-Dienst Standard

7 OGC Web Feature Service, Geometrie-Dienst Standard

8 OGC Web Coverage Service, Rasterdaten-Dienst Standard

grundlegende Sinnzusammenhänge, ohne die ihnen eine gemeinsame Grundlage fehlen würde, bin hin zum Beispiel oben im Jahr 1524 und bei Adam Rise.

REST - Technik mit Sinn verbinden

Die technische Lösung für einen Verweis im Internet, ist typischerweise ein Verweis in der Ausprägung einer URL. Ein häufig auftretendes technisches Problem ist die mangelnde und brüchige Verweis-Bindung (Linking, oder "Ver-Linkung"), die in meist wenig hilfreichen HTTP 404 (File not found) Statuscodes endet (siehe Abbildung 3).

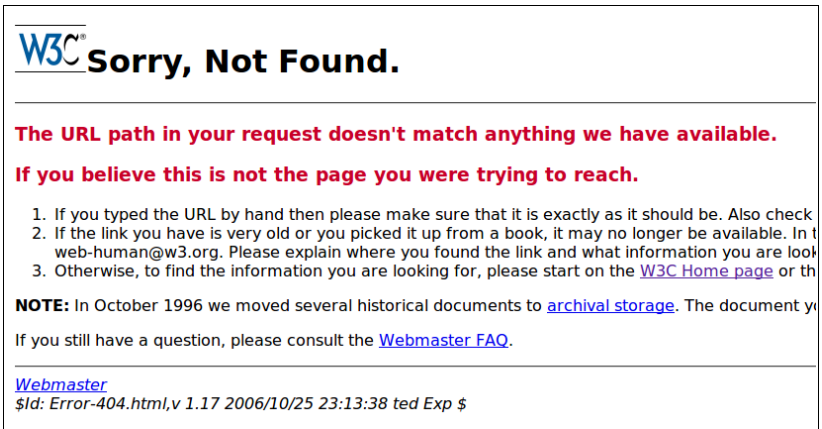

Abb. 3: Typische HTTP Statuscode 404 Fehlermeldung [6]

Die Lebensdauer von Verweisen (Adressen) kann sehr einfach durch die konsequente Nutzung der technischen Möglichkeiten erhöht werden. Das REST-Architekturparadigma beschreibt genau diesen Prozess. Das HTTP Applikationsprotokoll[9] des Internet bietet bereits alle erforderlichen Eigenschaften, z.B. den HTTP Statuscodes der 3er Serie "Umleitung" statt den der 4er Serie "File not found" (404)[10]. Leider sind die meisten Anbieter jedoch noch nicht für diese Technologie sensibilisiert, weshalb sie nicht oft zur Anwendung kommt. Der Statuscode "Umleitung" spezifiziert die neue Adresse der Ressource:

```
301 Moved Permanently

    Die angeforderte Ressource steht ab sofort unter der im „Loca-
    tion“-Header-Feld angegebenen Adresse bereit. Die alte Adresse
    ist nicht länger gültig.
```

Der Begriff "Ressource" ist hier im Kontext von REST und der Resource Oriented Architecture (ROA) sehr allgemein zu verstehen. Ein OCG WMS Dienst ist genauso wie ein PDF-Dokument aus der Perspektive des Internet nichts weiter als eine Ressource. Die Darstellung einer Karte, also das Ergebnis einer GetMap-Anfrage, ist eine Repräsentation dieser Ressource. Das Capabilities-Dokument des OCG WMS Dienstes ist nichts weiter als eine andere Repräsentation der gleichen Ressource.

Im Capabilities-Dokument eines Dienstes ist der Zugang zu allen anderen Repräsentationen enthalten, sei es eine Liste der Ebenen des Dienstes, die unterstützten Koordinatensysteme, Bildformate, Legenden-Elemente oder weitere Aspekte. Deshalb kommt dem Capabilities-Dokument eine ganz zentrale Bedeutung zu, vor allem der "Online-Resource". Ohne diese

9 World Wide Web Consortium (W3C); http://www.w3.org/hier_kommt_ein_404

10 Internet Engineering Task Force (IETF); RFD 2616; http://www.ietf.org/rfc/rfc2616.txt

Adresse ist der Dienst schlichtweg nicht erreichbar. Diese Repräsentation eines Kartendienstes ist aus Perspektive des weltweiten Netzwerkverbundes die wichtigste Metainformation überhaupt. Die hier vermerkte URL ist der seidene Faden an dem die gesamte weitere Nutzbarkeit des Dienstes hängt. Deshalb halten aktuelle Geoportale und Metadatenkataloge Kopien der Capabilities-Dokumente vor und überprüfen in regelmäßigen Abständen die Verfügbarkeit und die Aktualität der Kopie. Entdeckt der Monitoring-Prozess eine Änderung, schlägt er Alarm und benachrichtigt über eine Abo-Funktion[11] alle registrierten Benutzer, die dies wünschen (ob Mensch oder Maschine).

Eine Änderung dieser Adresse darf nicht unbemerkt erfolgen, sie muss sofort und überall, am besten automatisiert, berichtigt werden, z.B. über RSS[12]. Wenn der Dienst auch für Maschinen weiter auffindbar sein soll, darf die alte Adresse nicht einfach gelöscht werden, denn das würde nur in die Sackgasse 404 führen. Stattdessen muss der Webserver mit einem Statuscode 301 (permanently moved) antworten und im Location-Feld des Headers die neue Adresse angeben. Wenn das Capabilities-Dokument, und vor allem die Online-Resource URL stimmen, können alle weiteren Informationen hergeleitet werden, da sie in dem maschinenlesbaren XML Dokument in einer definierten Syntax hinterlegt sind.

Hierarchische Kataloge durch digitale Fundgruben ersetzen

Selbst wenn die technische Hürde "toter Links" überwunden wird, bleiben Geodaten häufig dennoch trotz umfangreicher, technischer Metadatenbeständen unauffindbar, da die Abbildung von Sinnzusammenhängen in herkömmlichen, hierarchisch strukturierten Katalogen nicht möglich ist. In den weiter oben gezeigten Beispielen sind die Informationen zu Entzerrung und Schattenaufhellung für sich genommen genauso wichtig, wie es zu wissen gilt, ob die Aufnahmeart der Verkehrsüberwachung amtlichen Charakter hat oder durch freiwillige Meldungen von Autofahrern erfolgt. Beide Informationen lassen sich aber schwerlich gemeinsam in einem Metadatenblatt oder Katalog abbilden, da sie strukturell unterschiedlichen Ontologien angehören. Deshalb sollten alle bereits verfügbaren digitalen Metadaten über Verweise (URL) an die Geodaten (Ressourcen) gebunden werden, um sie vollständig für die Suche nutzbar zu machen. So werden Metadaten in Wert gesetzt.

Weitere Einschränkungen bei der Bereitstellung von Geodatendiensten sind hausgemacht und teilweise nicht einmal technischer Natur, z.B. wenn die verwendete Software Unmengen von Daten über eine einzige Adresse bereitstellt. Ein typisches Beispiel sind Karten- und Geodatendienste, die mehrere hundert oder sogar tausend Ebenen enthalten. Diese lassen sich mit aktuellen, technisch durchaus der INSPIRE-Richtlinie genüge leistenden Metadaten zwar beschreiben, die dort hinterlegten Informationen sind aber ohne Modifikation nicht für die Weiterverarbeitung zu gebrauchen.

Anwender müssen sämtliche Metadaten **und** die Daten selbst mit ihren eigenen Ontologien durchsuchen können. Die Syntax der Suche wird durch die Technologie vorgegeben. Sie sollte möglichst einfach sein, da sie von den Anwendern erlernt werden muss . Ein viel zitier-

11 Geoportal-RLP: http://www.geoportal.rlp.de/mediawiki/index.php/Monitoring_abonnieren

12 Really Simple Syndication

tes Beispiel ist die einfache Google-Suche, der es reicht mit einem einzigen Textfeld (oft) sinnvolle Ergebnisse zu liefern. Die digitale Verfügbarkeit von Metainformationen kommt den Bedürfnissen der Geodaten-Anwender entgegen und entlastet den -Anbieter davon, sein eigenes Angebot nach für ihn unbekannten Kriterien zu strukturieren und zu beschreiben. Kommunikationsprozesse erfolgen vor allem zwischen Anwendern, schon allein weil deren Zahl viel höher ist als die der Anbieter. Alle die an diesem Prozessen teilnehmen, können sich in den Bereiche betätigen für die sie qualifiziert sind und die sie interessieren.

Im Web 2.0 gibt es eine Vielzahl technischer Möglichkeiten, um diese Kommunikation komfortabel und einfach zu gestalten. Ein aktuelles Beispiel, das bereits einen Teil dieser Möglichkeiten nutzt, ist das Web-basierte WMS Server Repository Geopole[13]. Hier können Dienste eingetragen, angezeigt und bewertet, sowie Informationen über neue Dienste und geänderte Metadaten per RSS[14] abonniert werden.

In der Übergangsphase, während die Datenanbieter noch lernen die Technologie korrekt einzusetzen, können Anwender Dienst-Ressourcen in Repositories[15] eintragen und mit Einträgen in Diskussionsforen, Wikis und originären Metadatenquellen verbinden. So entsteht eine heterogene Menge unsortierter Information unterschiedlicher Ontologien. Dadurch wird der Prozess der Organisation dynamisch und verlagert sich vom Anbieter zum Anwender. Der kategorisierende Prozess erfolgt jetzt bei der Suche, die mit einfacher, klar strukturierter Syntax aber hoch fachspezifischer Semantik erfolgt. Problem gelöst.

Kontakt zum Autor:

Arnulf Christl
Metaspatial
Heerstr. 162
0228 9768424
arnulf.christl@metaspatial.net

Literatur

[1] Kepa Korta, John Perry: Pragmatics, http://plato.stanford.edu/entries/pragmatics/, 2006

[2] Gruninger, M., Lee, J.: Ontology - applications and design. Comm. ACM 45(2), 39-41 (2002)

[3] Rise, Adams: Rechenbuch auff Linien und Ziphren in allerley Hanthierung, Annaberg, 1524

[4] Marine Metadata Interoperability http://marinemetadata.org/; 2010

[5] Ocean Data Standards http://www.oceandatastandards.org/; 2010

13 WMS Server Repository, http://www.geopole.org/

14 Really Simple Syndication, http://www.rssboard.org/rss-specification

15 Ablage, Fundgrube, Verwahrungsort: http://dict.leo.org/ende?lang=de&search=Repository