Datenqualität in freien Datenbanken

- Fassung 18.11.2010 -

Georg Verweyen

Konzernprofil und Größe.

- Die Deutsche Telekom ist eines der **weltweit führenden Dienstleistungsunternehmen** in der Telekommunikationsbranche.
- Unsere Stärke: Der Konzern bietet Produkte und Dienstleistungen aus den Bereichen **Festnetz, Mobilfunk, Internet** und **IPTV** für Privatkunden sowie **ICT-Lösungen** für Groß- und Geschäftskunden.
- Im Jahr 2009 betrug der Umsatz der Deutschen Telekom **64,6 Mrd. €**, mehr als die Hälfte des Umsatzes wurde 2009 im Ausland erwirtschaftet. Das bereinigte EBITDA lag bei **20,7 Mrd. €** und der Free Cash-Flow bei **7 Mrd. €**.
- Die **Gesamtzahl der Mitarbeiter** betrug zum 30. Juni 2010 über **251 000**.

Quelle: Geschäftsbericht 2009 und 1. HJ 2010

Angaben zur Person und dem Verfahren

- 20 Jahre bei der Telekom
- Zugehörig zum IT-Bereich, mit deutlicher Tendenz zum Data Warehouse-Umfeld, dort auch zwischenzeitlich für 6 Jahre im Marketing für die fachseitige Nutzbarkeit zuständig.
- Alle Beispiele sind in meiner Freizeit nur mit PHP und SQL-Standardmitteln ausgearbeitet worden (also kein Tooleinsatz).

Ziel des Foliensatzes

- Präsentationen zu Fragen der Datenqualität in firmeninternen Datenbeständen sind aus Datenschutzgründen selten erhältlich.
  - Präsentationen auf privater Datenmenge (z.B. Ahnenforschung) haben den Nachteil, dass sie nur bedingt nachvollziehbar sind.
- In diesem Vortrag werden häufig auftretende Probleme in drei größeren, frei verfügbaren Datenquellen beispielhaft präsentiert, um damit nachvollziehbare Erkenntnisse offen diskutieren zu können.
- Ziel des Vortrages: Toolhersteller können vergleichbare Aussagen über mögliche Verfahren und Performance dieser Analysen machen.

OpenLibrary.org

- Datenquelle: http://openlibrary.org/data/ol_dump_editions.txt.gz (4.7 GB)
- Ein großer Büchereikatalog
  - 20,7 Mio. Autoren
  - 23,8 Mio. Bücher
  - 30,8 Mio. Referenzen (ISBN10, ISBN13, LCCN, OCLC)
- Daten in JOSN-Struktur
- Objekt der Untersuchung:
  - ISBN10: 9 Ziffern mit Prüfziffer ([1]) an 10. Stelle (kann ein X sein)
  - ISBN13: 12 Ziffern mit Prüfziffer ([2]) an 13. Stelle

([1]) Gewichtungsfaktoren (1, 2, 3, 4, 5, 6, 7, 8, 9 Modulo 11) Prüfziffer 10 wird als X dargestellt
([2]) Gewichtungsfaktoren (1, 3,1, 3, 1, 3,1, 3, 1, 3, 1, 3 Modulo 10)

Aussage zu DQ Openlibrary

- Ein Standardprofiling der Daten ergibt

Länge 10	64,1%
Länge 13	35,6%

  und teilweise erkennt man Preise in den Feldern. Diese Prozentangaben sind jedoch irreführend.

- Nach den Referenztypen getrennt ergibt sich für die korrekte Länge

ISBN10	99,63%
ISBN13	99,96%

- davon sind die Prüfziffern fehlerhaft:

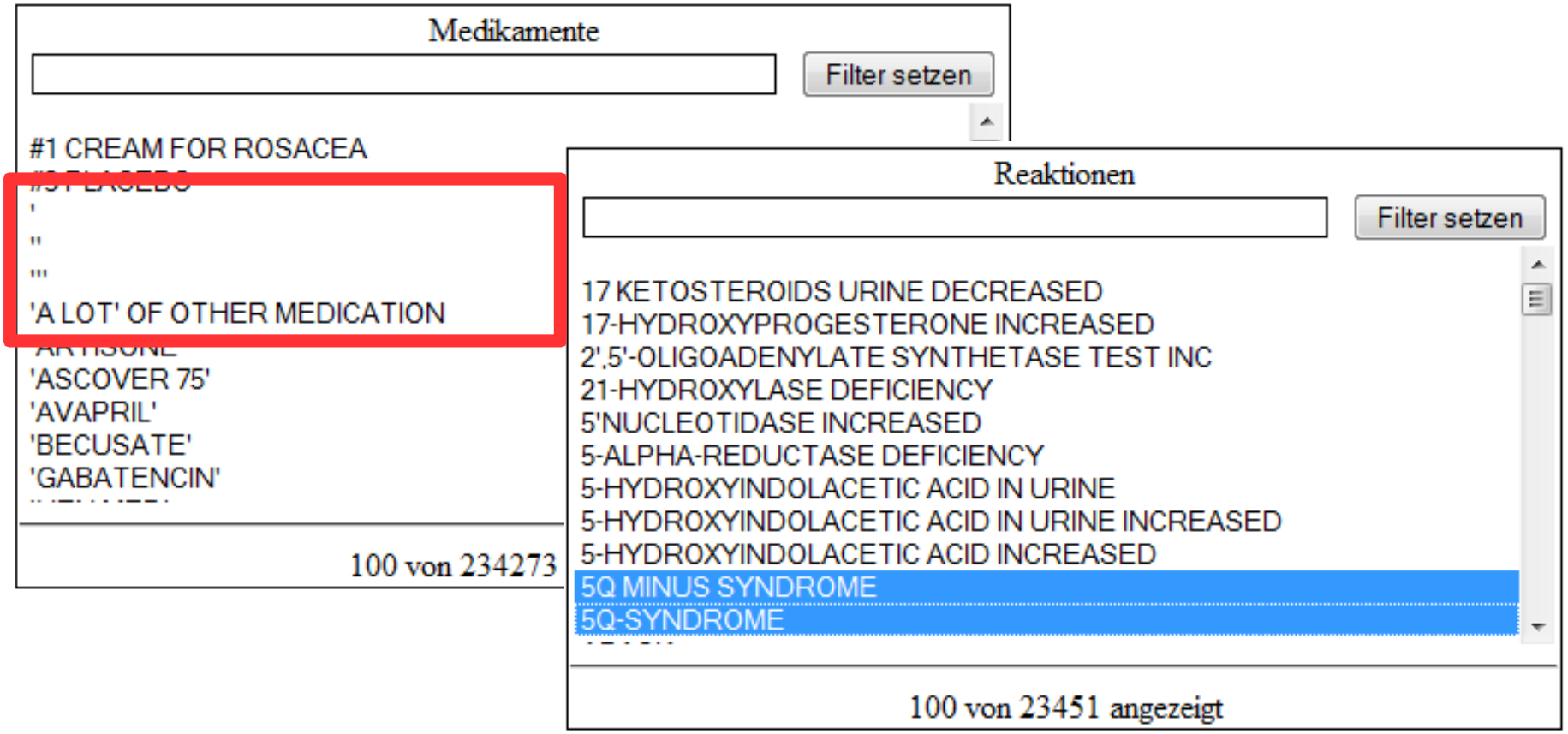
ISBN10	3,0 ‰
ISBN13	0,8 ‰

FDA
Food and Drug Administration, USA

- Datenquelle: http://www.fda.gov ([1])
- Quartalsweise Sammlung von potentiellen Nebenwirkungen von Medikamenten
  - 81,3 Tausend Vorfälle mit 112 Tausend Detailberichte
  - 23,6 Tausend Medikamente
  - 30,3 Tausend Reaktionen
- Daten in <XML>-ähnlicher-Struktur
- Objekt der Untersuchung:
  - Schreibweise der unterschiedlichen Medikamente (234.273, alle Quartale)

([1]) http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm

FDA
Probleme bei Freitext Eingaben



- Allein durch Leerzeichen wird die Anzahl der Medikamente um 1,1% erhöht.

FDA
Toolwunsch Erkennbarkeit von Strukturen

Bei der Pattern-Analyse sollte die Folge von gleichen Typen zusammengefasst werden können, um grundlegende Strukturen besser erkennen zu können.

Auch wenn nur wenige mit Medikamenten Eingaben zu haben, sind häufig E-Mailadressen vorhanden.

VITAMIN E	/00110501/	ALAE/Z/
VITAMIN E	/00110501/(TOCOPHEROL)	ALAE/Z/L(A)
VITAMIN E	/GFR/	ALAE/A/
VITAMIN E	/001105/	ALAE/Z/
VITAMIN E	/00110501/	ALAE/Z/
VITAMIN E	/001105/	ALAE/Z/
VITAMINE E	/001105/	ALAE/Z/

E-Mail Pattern

x@x.x

x.x@x.x

x.x.x@x.x

x@x.x.x

x.x@x.x.x

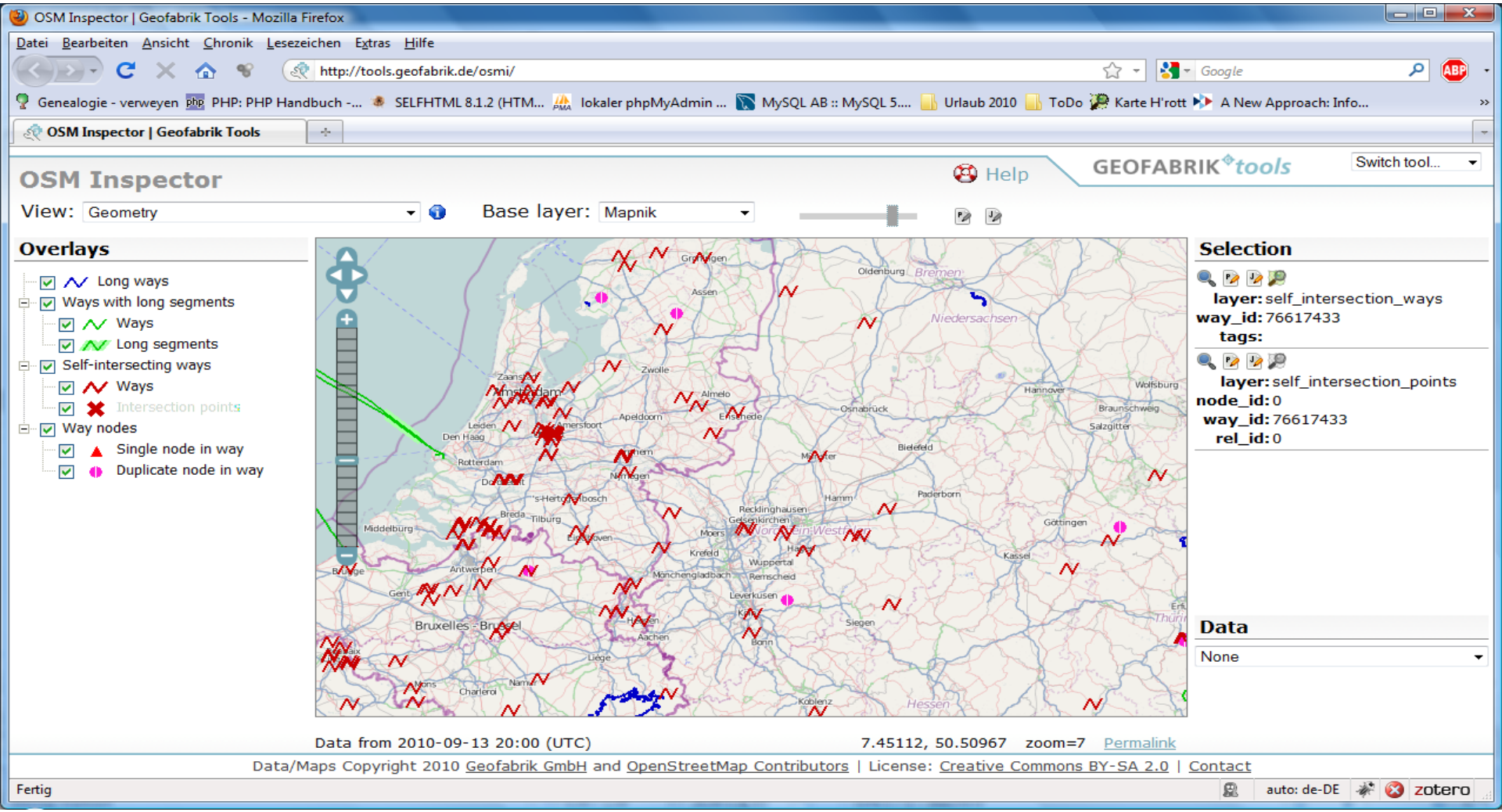
OpenStreetMap.org

Datenquelle: http://download.geofabrik.de/osm/ ([1])

- Daten zur Kartendarstellung ([2])
  - 774 Millionen Knoten mit Geokoordinaten
  - 63,8 Millionen Wege
  - 0,7 Millionen Relationen (Zusammenfassung von Wegen und Knoten)
  - 64,9 Millionen Beschreibungswerte (Tags)
- Daten in <XML> Struktur
- Objekt der Untersuchung
  - Vollständigkeit, Wertekombinationen, Werteausprägung und ...

([1]) eine mögliche Quelle mit unterschiedlichen Ausschnitten, Details: siehe wiki.openstreetmap.org
([2]) Stand 15. September 2010

OpenStreetMap.org
visuelle Fehlerkontrolle



Aussagen zur DQ in OpenStreetMap.org

- Laut der Webseite des ADAC gab es im Januar 2010 14.410 Tankstellen ([1]). Im deutschen Datenbestand von OSM waren am 23.04.2010 13.087 Tankstellen (ohne diese Einschränkung) erfasst (Erfassungsquote 90,8%), Anzahl stark steigend.
- Derzeit gibt es circa 20.100 unterschiedliche Beschreibungen, etwa 7.000 (35%) existieren nur einmal im Datenbestand

([1]) Es wurden nur Straßentankstellen erfasst; ohne Autobahntankstellen und reine Biodiesel- bzw. Erdgas-tankstellen

Fazit

- Vorgestellt wurden drei größere, frei verfügbare Datenbanken, die
  - größere Datenmenge umfassen und
  - repräsentative Datenqualitätsprobleme besitzen,
  - deren Ursachen auf der Verarbeitungskette beruhen.
- Mit diesen oder ähnlichen frei verfügbaren Datenbank können Toolhersteller vergleichbare Aussagen
  - über mögliche Verfahren und
  - die Performance dieser Analysen machen.

Zusammenfassung

Anbieter in der Workshop-Reihe werden sich dem Thema FDA widmen.

Fragen?

URL‘s:

- http://www.t-mobile.de
- http://www.familieverweyen.de