Semistrukturierte Daten

Einführung

Stefan Woltran
Emanuel Sallinger

Institut für Informationssysteme
Technische Universität Wien

Sommersemester 2014

Was sind Semistrukturierte Daten?

Strukturierte Daten

- bezieht sich meistens auf das relationale Datenmodell
- Strukturierung in Tabellen (Relationen)
- jede Zeile einer Tabelle hat die gleichen Attribute

Semistrukturierte Daten

- Daten nicht nach dem relationalen Datenmodell strukturiert
- Schema nicht notwendigerweise vorhanden ("selbstbeschreibend")

Flexible Struktur

- Semistrukturierte Daten sind geeignet für flexible und unregelmäßige Daten
  - wenn eine relationale Datenbank viele Nullwerte hätte
  - wenn Daten unterschiedliche Typen haben können
- Gut geeignetes Datenmodell:
  - Baum
  - Graph
- Relationale Daten lassen sich als Bäume oder Graphen leicht darstellen

Selbstbeschreibende Daten

- Schema bei relationalen Datenbanken:
  - zuerst wird ein Schema definiert (Struktur und Datentypen)
  - erst dann werden Daten eingefügt
- Schema bei semistrukturierten Daten kann:
  - nicht vorhanden,
  - nicht bekannt oder
  - ständig evolvierend sein
- Selbstbeschreibende Daten
  - Daten werden mit Beschreibung annotiert
  - Vorteil: interoperabel, erweiterbar
  - Nachteil: erhöhter Speicherplatzbedarf

Dokumente vs. Daten

- Dokumente
  - z.B. Präsentationsformate (HTML)
  - reiner Text ist problematisch für automatische Verarbeitung von Inhalten
  - Sichtbarmachen der Struktur (mittels Markup) hilfreich
- Daten
  - z.B. relationale Datenbanken
  - starre Struktur, fixes Schema
  - selbstbeschreibende Daten (mittels Markup) geben mehr Flexibilität
- Semistrukturierte Daten
  - vereinigt beide Sichtweisen
  - Dokument- und Datensicht wird von verschiedenen Standards für semistrukturierte Daten in unterschiedlicher Ausprägung unterstützt

Semistrukturierte Daten und XML

- XML ist ein guter Repräsentant für semistrukturierte Daten:
  - Familie von Standards (Namespaces, Schemasprachen, Abfragesprachen)
  - starke Verbreitung im Enterprise Bereich
  - gute Unterstützung durch Tools
- XML ist nicht die einzige Sprache für semistrukturierte Daten:
  - JSON als Format für Datenaustausch bei Webanwendungen
  - YAML als Format für Konfiguration, Logs, etc.
  - ...

Semistrukturierte Daten jenseits von Bäumen

Semistrukturierte Daten sind nicht auf Bäume beschränkt:

- z.B. Graphdaten im Semantic Web (RDF)

Wann sind Daten "semistrukturiert"?

- Key-Value Daten
- Relationale Daten (Tabellen) ohne Schema