Semantische Datenintegration

Integration von Daten

Jakob Voß

2017-04-01

Übersicht

- Definitionen
  - Informationsintegration vs. (semantische) Datenintegration
- Anwendungen
  - Integrierte Informationssysteme
  - Übungsbeispiel
- Herausforderungen
  - Verteilung
  - Autonomie
  - Heterogenität

Definitionen

Informationsintegration

Informationsintegration ist die Zusammenführung von Daten und Content verschiedener Quellen zu einer einheitlichen Informationsmenge.

Informationsintegration

> *Informationsintegration ist die **korrekte, vollständige** und **effiziente** Zusammenführung von Daten und Content verschiedener, **heterogener** Quellen zu einer einheitlichen und **strukturierten** Informationsmenge zur **effektiven Interpretation** durch Nutzer und Anwendungen.*

(Naumann und Leser 2006)

Datenintegration

Zusammenführung von Daten verschiedener Quellen

Semantische Datenintegration

Semantische Datenintegration ist die **sinnvolle** Zusammenführung von **uneinheitlichen** Daten verschiedener Quellen.

Anwendungen

Einfache Datenintegration

- Einmaliges und/oder manuelles Zusammenführen von Daten
- Beispiele
  - Migration auf neue Systeme
  - Zusammenlegung von Datenquellen oder Organisationen
- Flexibel und manchmal notwendig
- Aufwändig und skaliert nicht

Integriertes Informationssystem

- Informationssystem: Umfassender Begriff für Computersysteme zum Sammeln, Speichern, Verarbeiten und gezielten Wiedergeben von Informationen

Integriertes Informationssystem

- Verhält sich nach außen wie *eine* Datenquelle
- Fasst intern verschiedene, heterogene Datenquellen zusammen
- Details siehe Informatik Naumann und Leser (2006) u.A.

Beispiele für integrierende Informationssysteme

- ...
- ...
- ...

Beispiele für integrierende Informationssysteme

- Metasuchmaschine
- Data Warehouse
- Mashup
- ...

Übung: Mashup von Bibliotheksdaten

- Aufgabe
  - Gegeben eine Bibliothek (ISIL) und Buch (PPN)
  - Name der Bibliothek und Signaturen der Exemplare
- Programmiersprache
  - Google Sheets
- Datenquellen
  - http://sigel.staatsbibliothek-berlin.de/suche/linked-data-service/
  - http://daia.gbv.de/

Übung: Mashup von Bibliotheksdaten

	A	B
1	ISIL	DE-18
2	PPN	508902037
3	ID	opac-de-18:ppn:508902037

Übung: Mashup von Bibliotheksdaten

	A	B
1	ISIL	DE-18
2	PPN	508902037
3	ID	`="opac-" & LOWER(B1) & ":ppn:" & B2`

Übung: Mashup von Bibliotheksdaten

```
="http://daia.gbv.de/?format=xml&id=" & B3
="http://ld.zdb-services.de/data/organisations/" & B1 & ".html"

=IMPORTXML(B6, "/*/*/*/*[local-name()='label']")
=IMPORTXML(C4,"//h2")`
```

Übung: Mashup von Bibliotheksdaten

	A	B
1	ISIL	DE-18
2	PPN	508902037
3	ID	opac-de-18:ppn:508902037
4		http://ld.zdb-services.de/data/organisations/DE-18.html
5		Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky
6		http://daia.gbv.de/?format=xml&id=opac-de-18:ppn:508902037
7		A 2007/1082
8		9/59706
9		D LES
10		Präsenzexemplar D LES 39615
11		ST 270 L628 59706
12		JL D LES 39612

Reflexion

- Zusammenführung verschiedener Quellen
- Identifier helfen ungemein
- Abfragesprachen allerorten

Herausforderungen

Uneinheitliche Daten verschiedener Quellen

- Eigene, gemeinsame und fremde Daten
- Verschiedenste Datenbanken
- Dateien (Text, Tabellen. . .) verschiedenster Formate
- Webseiten und -Dienste
- Beliebige Software
- . . .

Schwierigkeiten

- Verteilte Datenquellen (Verteilung)
- Datenquellen sind unabhängig (Autonomie)
- Daten sind uneinheitlich (Heterogenität)

Verteilung

- Unterschiedliche Programme
- Unterschiedliche Server
- Unterschiedliche physikalische Orte
- Unterschiedliche Organisationen

Kommunikation trotzdem möglich!

Verteilung: Vor- und Nachteile

- Vorteile
  - Ausfallsicherheit
  - Austausch- und Teilbarkeit
  - Autonomie
- Nachteile
  - Komplexität
  - Autonomie und Heterogenität

Autonomie

- Datenquellen sind unabhängig voneinander
- Keine Instanz kontrolliert alle Teile
- Details der Implementierung können sich jederzeit ändern!
- Datenquellen sind nicht unbedingt darauf ausgelegt, einfach integriert zu werden (u.A. keine stabilen APIs)

Autonomie: Beispiele

Schauen wir uns einige Datenquellen an, bezüglich

- Organisation und Betrieb
- Zugriffsmöglichkeiten (Anfrage, Übertragung, Format. . .)
- Einflussmöglichkeiten

Autonomie: Vor- und Nachteile

- Vorteile
  - Flexibilität & Gestaltungsspielraum
- Nachteile
  - Unkoordinierte Datenquellen
  - jede(r) macht nur seins
  - Führt zu Heterogenität

Föderiertes Informationssysteme

- Autonome Bestandteile
- Föderation durch Absprachen
- Koordination möglich
- Absprachen müssen exakt festgelegt und überprüft werden (gemeinsame Standards)

Beispiel: Das WWW

Heterogenität

Einteilung nach (Naumann und Leser 2006)

- syntaktische/technische Heterogenität
- strukturelle Heterogenität
- semantische Heterogenität

Syntaktische/technische Heterogenität

- Unterschiedliche Hardware(möglichkeiten)
- Unterschiedliche Software
- Unterschiedliche Zugriffsmethoden
  - Schnittstellen (Z39.50, REST, HTTP-irgendwas...)
  - Anfragesprachen (SQL, SPARQL, Web-Formular...)
  - Anfrageparameter (Variablen und Felder)

Strukturelle Heterogenität

- Unterschiedliche Formate und Strukturierungssprachen
  - SQL, XML, JSON, RDF, MARC...
- Unterschiedliche Umsetzung der Modellierung

Strukturelle Heterogenität: Beispiel

```
{
  "author": "Emma Goldman",
  "year": "2014",
  "isbn": "978-3-89401-810-8"
}
```

```
{
  "author": [ "Goldman, Emma" ],
  "date": "2014",
  "isbn": "9783894018108"
}
```

Strukturelle Heterogenität: Lösungen

- Konvertierung zwischen Formaten
- Mapping von Schemas
- Genauere Anwendungsregeln

Semantische Heterogenität

- Unterschiedliche Bedeutung
- Unterschiedliche Modelle
- Unterschiedliche Grundannahmen
- Abgrenzung zu struktureller Heterogenität ist oft Teil des Problems

Semantische Heterogeniät auf Schemaebene

- Ist mit “author” das gleiche gemeint
  oder sind es homonyme Konzepte?
- Sind “date” und “year” Homonyme Konzepte?
- Spielt die Formatierung der ISBN eine Rolle?
- Unterschiedliche Einheiten (z.B. $ vs. €) und Sprachen

Semantische Heterogeniät auf Datenebene

- Identität: Wann beziehen sich Datensätze oder Felder auf das gleiche Objekt?
  - $\rightarrow$ Objektidentifikation und Duplikaterkennung
- Widersprüchliche Angaben (Datenkonflikte)
  - Zwei Datenquellen widersprechen sich
  - Zwei Datenquellen ergänzen sich

Ursachen für Datenkonflikte

- Uneinheitliche Ansetzungen/Schreibvarianten
- (Tipp)fehler
- Veraltete Daten
- Fehlende Angaben
  - Closed World Assumption
  - Open World Assumption

Zusammenfassung der Herausforderungen

- Datenquellen deren Zusammenführung sinnvoll wäre sind oft verteilt und autonom
- Datenquellen werden damit unkoordiniert und uneinheitlich
- Dies führt zu Heterogenität
- Herausforderung semantischer Datenintegration: **Überbrückung der Heterogenität**

Zusammenfassung

- Integriertes Informationssystem
  - Einheitliche Sicht über verschiedene Quellen

Literatur

Quellen dieser Folien: https://github.com/hshdb/MWM-317-02/

Naumann, Felix, und Ulf Leser. 2006. *Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen*. Heidelberg: dpunkt-Verlag.