

# Data-Warehouse-Praktikum

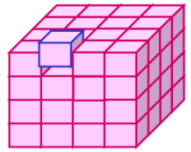
Einführungsveranstaltung WS 2011/12

Lars Kolb

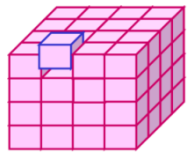
Abteilung Datenbanken, Universität Leipzig

<http://dbs.uni-leipzig.de>

# Organisatorisches

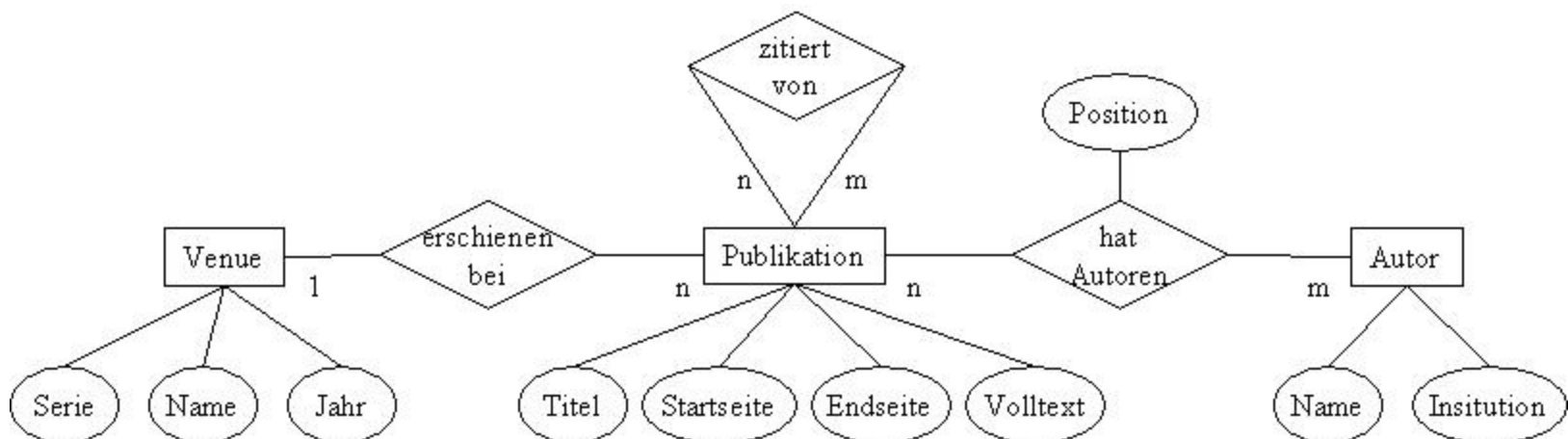


- Ziel: Realisierung eines "typischen" DWH-Projekts
  - Kennenlernen der "echten, praktischen" DWH-Probleme
- Zielgruppe
  - Informatik-Studenten (Master, Diplom)
  - Interessierte
- Kenntnisse
  - Vorlesung "Data Warehousing" nötig
  - Vorlesung "Datenintegration" hilfreich
  - Skripte zum Selbststudium/Nacharbeiten im Netz
- Ablauf
  - Gruppenarbeit mit 2-3 Studenten pro Gruppe
  - Bearbeitung von 3 Aufgaben → jeweils Testat → Schein
- Aufgabenstellung und Informationen
  - <http://dbs.uni-leipzig.de/de/stud/2011ws/dwhprak>

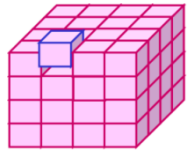


# Szenario: Zitierungsanalyse

- Innerhalb wissenschaftlicher Arbeiten werden andere Arbeiten zitiert
- Anzahl der Zitierungen charakterisiert wissenschaftlichen Einfluss
  - Wie häufig wird Publikation X zitiert?
  - Wie häufig werden Publikationen des Venues (= Konferenz oder Journal) Y im Durchschnitt zitiert?
  - Wie ist die durchschnittliche Zitierungszahl von Autor Z?

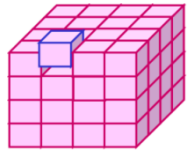


# Datenquellen



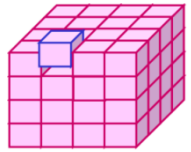
- DBLP Bibliography
  - manuelle gepflegte Website, die komplette Listen verschiedener Venues aus dem Informatik-Bereich enthält.
- ACM Digital Library
  - Portal der Association for Computing Machinery
  - enthält ebenfalls komplette Listen verschiedener Venues
- Google Scholar
  - Suchmaschine für wissenschaftliche Publikationen
- Relevante Teilmenge der Daten steht als CSV- und XML-Dateien zur Verfügung

# Aufgaben: Inhaltlich



- Datenimport
  - Import der XML- und CSV-Dateien
  - Relationale Speicherung der Daten
- Data Cleaning
  - Objektkonsolidierung: Erkennen gleicher Publikationen in verschiedenen (oder gleichen) Datenquellen
  - Datennormalisierung: Normalisierung der Institutionsnamen
  - Ableitung neuer Daten: Identifikation von Selbstzitierungen
- Cube-Erstellung, OLAP und Data Mining
  - Star-Schema-Erstellung und Datenimport
  - OLAP-Analyse, MDX-Anfragen
  - Data Mining: Assoziationsregeln zur Bestimmung "ähnlicher Venues"

# Aufgaben: Organisatorisch



- Realisierung mittels "SQL Server Business Intelligence Development Studio"
  - "Drag&Drop"-Workflow-Erstellung (keine Programmierung)
  - Verfügbar auf Windows-Rechnern im Pool
    - Remote: [wserv1.informatik.uni-leipzig.de](http://wserv1.informatik.uni-leipzig.de)
    - Login: Bekanntgabe nach Anmeldung/Einführungsveranstaltung
  - Client-Anwendung für zentralen Datenbankserver (SQL Server 2008 auf dbserv1)
- Jeder Aufgabe ist ein Tutorial zugeordnet
  - Beschreibung der Aufgabe
  - Grundlegende Vorgehensweise (incl. Screenshots)
  - Weitere Hinweise
- Software-Ergebnis sind ausführbare Projekte, welche im Testat ausgeführt/begutachtet werden
  - Terminabsprache per E-Mail, Deadlines siehe Webseite

