Web Data Integration

Introduction and Course Outline
Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web-based Systems
  - Data and Web Mining
  - Identity Resolution
  - Data Quality and Fusion
- Room: B6, 26 - B1.15
- Consultation: Wednesday 13:30-14:30
- eMail: chris@informatik.uni-mannheim.de
Hallo

- M. Sc. Wi-Inf. Oliver Lehmberg
- Graduate Research Associate
- Research Interests:
  - Data and Web Mining
  - Network Analysis
  - Web Data Integration
- Room: B6, 26, C 1.04
- eMail: oli@informatik.uni-mannheim.de

- Will teach the practical parts and supervise the student projects.
Introduction and Course Outline

1. What is Data Integration?
2. Application Areas
3. Types of Heterogeneity
4. The Data Integration Process
5. Data Integration Architectures
6. The Data Integration Software Market
7. Course Outline and Organization
1. What is Data Integration?

- Databases are great: They let us manage huge amounts of data.
  1. Assuming you’ve put it all into a single schema.
  2. Assuming the database doesn’t contain duplicate records.
  3. Assuming that there are no data conflicts.

- In reality, applications often need to work with data from multiple independently created data sources.
  1. Different sources use different data models.
  2. Different sources use different schemata.
  3. Different sources describe the same real-world entity.
  4. Different sources provide conflicting data about a single entity.
  5. Different sources provide different limited query interfaces to their data.
What is Data Integration?

- The integrated data set should:
  1. Correctly and completely represent the content of all data sources.
  2. Use a single data model and a single schema.
  3. Only contain a single representation of every real-world entity.
  4. Not contain any conflicting data about single entities.

- To achieve this, data integration needs to resolve various types of heterogeneity that exist between data sources.

Data integration is the process of consolidating data from a set of heterogeneous data sources into a single uniform data set or view on the data.
2. Application Areas

1. Business
2. Science
3. Government
4. The Web
5. .... pretty much every application area
Application Area: Business

Enterprise Databases

Legacy Databases

Services and Applications

Oracle estimate: 50% of all IT $$$ are spent here!

- CRM
- SCM
- Business Intelligence
- Company Mergers
- ...
Application Area: Science

Hundreds of biomedical data sources available; growing rapidly!
Law enforcement agencies integrate data from various sources in order to identify suspects.

- Cell phone calls
- Location data
- Online profiles (Facebook)
- Web browsing behavior
- Credit card transactions
- Intelligence from other agencies
- ...
Application Area: Data Journalism

- Government data is increasingly published under open licenses on the Web.
- Journalists discover stories by combining data from different sources.

EU subsidies
- received for renovating a ship
- received for scraping the same ship

Members of parliament
- donations/membership in supervisory boards
- voting behavior
Application Area: Online Shopping
Comparison Shopping

The Unofficial Harry Potter Cookbook: From Cauldron Cakes to Knickerbocker Glory--More Than 150 Magical Recipes for Muggles and Wizards [Book]

$3 online


Bangers and mash with Harry, Ron, and Hermione in the Hogwarts dining hall. A proper cuppa tea and rock cakes in Hagrid’s hut. Cauldron cakes and pumpkin juice on the Hogwarts Express. With this cookbook, dining a la Hogwarts is as easy as Banoffee Pie! With more than 150 easy-to-make ... more »

Online stores  Reviews  Details

Sellers  Seller Rating  Details  Base Price  Total Price

MovieMars.com  ★★★★☆ (42)  Free shipping  $20.92

ValoreBooks.com  No rating  No tax  $3.24 + $3.95 shipping  $7.19

Overstock.com  ★★★★☆ (5,086)  $13.93

Sponsored  ‡
The Deep Web is accessible via HTML Forms
Structured Data on the Web

More and more Websites

- semantically markup the content of their HTML pages
- publish structured data in addition to HTML pages
We distinguish five types of heterogeneity:

1. Technical Heterogeneity
2. Syntactical Heterogeneity
3. Data Model Heterogeneity
4. Structural Heterogeneity
5. Semantic Heterogeneity

The goal of data integration is to bridge all these types of heterogeneity.

Data source autonomy is the main reason for heterogeneity:
- Data sources independently decide how to store things and how to provide access
- Agreeing on standards partly reduces heterogeneity
### Technical Heterogeneity

Technical heterogeneity comprises all differences in the means to access data, not the data itself.

<table>
<thead>
<tr>
<th>Level</th>
<th>Possibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication Protocol</td>
<td>HTTP, ODBC/JDBC, SOAP</td>
</tr>
<tr>
<td>Data Exchange Format</td>
<td>XML, JSON, CSV, RDF, HTML, binary data</td>
</tr>
<tr>
<td>Query Language</td>
<td>Full query language: SQL, SPARQL</td>
</tr>
<tr>
<td></td>
<td>Canned queries: Web Services, Web Forms</td>
</tr>
<tr>
<td></td>
<td>Download of complete data set dumps</td>
</tr>
<tr>
<td>Additional Restrictions</td>
<td>Number of queries</td>
</tr>
<tr>
<td></td>
<td>Cost per query / data set</td>
</tr>
<tr>
<td></td>
<td>Access rights</td>
</tr>
</tbody>
</table>
**Syntactical Heterogeneity**

Syntactical heterogeneity comprises all differences in the encoding of values.

<table>
<thead>
<tr>
<th>Level</th>
<th>Possibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character format</td>
<td>ASCII versus Unicode</td>
</tr>
<tr>
<td>Number format</td>
<td>Little endian versus big endian</td>
</tr>
<tr>
<td>Delimiter format</td>
<td>Tab-delimited versus Comma-separated values</td>
</tr>
</tbody>
</table>

Syntactical heterogeneity does not comprise

- Synonymous values
  - 1GB versus 1000MB ➔ Semantic heterogeneity
- Structural differences
  - First name: Chris, last name: Bizer versus name: Chris Bizer ➔ Structural heterogeneity
Data Model Heterogeneity

Data model heterogeneity comprises differences in the **data model** that is used to represent data.

Data Models:

1. Relational data model
2. XML data model
3. Object-oriented data model
4. RDF graph data model
Structural heterogeneity comprises differences in the way different schemata represent the same part of reality.

1. Alternative Modeling
   - Relation vs. Attribute
   - Attribut vs. Value
   - Relation vs. Value

2. Normalized vs. Denormalized

3. Nested vs. Foreign Key Relationship
Example: Alternative Modelling

Man( Id, Firstname, Surname)
Woman( Id, Firstname, Surname)

Relation vs. Attribute

Relation vs. Value

Person( Id, Firstname, Surname, Sex)

Person( Id, Firstname, Surname, Male, Female)

Attribute vs. Value
Semantic Heterogeneity

Semantic heterogeneity comprises differences concerning the meaning of data and schema elements.

1. Naming Conflicts
   - Synonyms, homonyms, slightly deviating concepts

2. Object Identity / Duplicates
   - Multiple data sources as well as multiple records within one data source may describe the same real-world entity.
   - Which “Franz Müller” does a record describe?

3. Data Conflicts
   - Conflicting data about the same real-world entity in different data sources as well as within different records in the same data source.
Semantic Heterogeneity: Synonyms

1. Synonymous schema element names:

   DB1:
   Employee( Id, FirstName, Name, Male, Female)

   DB2:
   Person( Id, FirstName, Surname, Sex)

2. Synonymous attribute values:
   - Different value coding schemas: Manager vs. 2
   - Different spellings / abbreviations: Kantstr. vs. Kantstrasse vs. Kant Str.
   - Different units of measurement: 1 GB vs. 1000 MB

Different words having the same meaning.
Semantic Heterogeneity: Homonyms

- Reason: Different people (in different situations) associate different meanings with the same word.

- Examples:

  DB1:
  Employee( Id, Name, Salary, m, f, Title)

  DB2:
  Person( Id, Name, Salary, Sex, Title)

  USD

  Secretary, Engineer, Manager, etc.

  Euro

  Mr., Mrs., Dr., Prof. Dr., ...
Problem: Precision of Concept Definitions

Business question: How many employees has IBM?

- Definition of Employee:
  - Temporary employees?
  - Students writing master theses?
  - External consultants?
  - Positions in organization chart or currently employed people?

- Definition of IBM
  - Which global region? Which business unit?
  - Include companies that are partly owned by IBM?

- Which point in time?
- How to count people that work part-time?
Problem: The same real-world entity is often represented
  • within multiple data sources.
  • by multiple records within the same data base.

- Relevant for: Product data, customer data, scientific data, …
- Business question: How much hardware did we sell to the University of Mannheim?
- Problem: CRM database likely contains multiple records referring to the university itself as well as the different faculties/professors.
- Reasons for duplicates in the same data base:
  • Different people entered data without identity checks
  • Same entity observed several times
  • No consistent global IDs in input data (ISBN, IBAN, URL, EAN, …)
Problem: Two duplicate records contain different values for the same attribute.

Reasons for data conflicts
1. **Errors**: Typos and other errors when data is entered.
2. **Outdated data**: One source/record is older than the other one.
3. **Disagreement**: Different sources actually disagree on the correct value / the truth.
4. The Data Integration Process

- Data Collection
- Schema Mapping
- Data Translation
- Identity Resolution
- Data Quality Assessment
- Data Fusion
4.1 Data Collection

Goal: Resolve technical and data model heterogeneity so that data from all sources can be accessed / gathered and represented in the same data model.

- Using middleware libraries that provide
  - different communication protocols (HTTP, ODBC, …)
  - readers for different data exchange formats (XML, RDF, JSON, …)
  - for querying remote data sources using different query languages (SQL, SPARQL, …)
  - for crawling remote data sources (HTML pages, Web APIs, Linked Data)
  - for translating data between different data models (XML-2-Relational, …)
Goal: Resolve structural and schema-related semantic heterogeneity by
1. finding correspondences between the elements of the different schemata.
2. translate data to a single target schema based on these correspondences.

ARTICLE
- artPK
- title
- pages

AUTHOR
- artFK
- name

PUBLICATION
- pubID
- title
- date
- author

Source Schema
Correspondences
Transformation Query
Target Schema
Target Data
Example: Defining Correspondences
4.3 Identity Resolution

Goal: Resolve semantic heterogeneity by identifying all records in all data sources that describe the same real-world entity.

Other names for the task:
- Duplicate Detection, Record Linkage, Entity Matching

Basic Approach:
1. Compare records using a combination of different similarity metrics
2. If similarity is above threshold → Consider records to describe the same real-world entity

<table>
<thead>
<tr>
<th>DB1</th>
<th>CID1243</th>
<th>Chris Miller</th>
<th>12/20/1982</th>
<th>Bardon Street, Melville</th>
<th>32 sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB2</td>
<td>34</td>
<td>Christian Miller</td>
<td>2/20/1982</td>
<td>7 Bardon St., Melville</td>
<td>24 sales</td>
</tr>
<tr>
<td>DB3</td>
<td>427859</td>
<td>Chris Miller</td>
<td>12/14/1973</td>
<td>7 Bardon St., Madison</td>
<td>13 sales</td>
</tr>
</tbody>
</table>
Example: Combining different Similarity Metrics
4.4 Data Fusion

Goal: Resolve data conflicts by combining attribute values of duplicate records into a single consolidated description of an entity.

**Basic Approach:**

1. Assess the **quality** of data sources / records / values
   - Quality dimensions: timeliness, reputation of source, ...
2. Apply a **conflict resolution function** to choose most promising values or to correct values.
   - Example functions: highest estimated quality, voting, average, ...

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Date</th>
<th>Address</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB1</td>
<td>CID1243</td>
<td>12/20/1982</td>
<td>Bardon Street, Melville</td>
<td>32</td>
</tr>
<tr>
<td>DB2</td>
<td>34</td>
<td>2/20/1982</td>
<td>7 Bardon St., Melville</td>
<td>24</td>
</tr>
<tr>
<td>Fused</td>
<td>Christian Miller</td>
<td>12/20/1982</td>
<td>7 Bardon Street, Melville</td>
<td>56</td>
</tr>
</tbody>
</table>
5. Data Integration Architectures

1. **Materialized Integration**
   - Integrate sources by bringing the data into a single physical database (data warehouse).

2. **Virtual Integration**
   - Leave the data at the sources and access it at query time via wrappers (integrated view).

3. **Numerous intermediate architectures**
## Materialized versus Virtual Integration

<table>
<thead>
<tr>
<th></th>
<th>Materialized Integration</th>
<th>Virtual Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data currency</td>
<td>Low (regular updates)</td>
<td>High (always current)</td>
</tr>
<tr>
<td>Storage requirements</td>
<td>High (copy all data locally)</td>
<td>Low (data remains in sources)</td>
</tr>
<tr>
<td>Query processing time</td>
<td>Low (local query processing)</td>
<td>High (slow network traffic)</td>
</tr>
<tr>
<td>System Complexity</td>
<td>Low (like normal DB)</td>
<td>High (planning of distributed queries)</td>
</tr>
<tr>
<td>Query Expressiveness</td>
<td>High (like normal DB)</td>
<td>Low (as sources might be restricted)</td>
</tr>
<tr>
<td>Workload on data source</td>
<td>Can be planned</td>
<td>Hard to plan</td>
</tr>
<tr>
<td>Identity Resolution / Data Fusion</td>
<td>possible</td>
<td>not possible (too slow)</td>
</tr>
</tbody>
</table>

- **Rule of thumb:** Virtual integration not applicable
  - if 5+ data sources need to be joined.
  - identity resolution and data fusion are important.

- This course illustrates data integration through the **materialized architecture**.
6. The Data Integration Software Market

- Market size 2013:
  2.3 billion US$ (growth: 9.4%)
- Tools for specific tasks
  - Altova Map Force
- Comprehensive solutions covering the complete data integration process
  - Informatica Plattform
  - IBM InfoSphere Information Server
  - SAP Data Services, SAP Replication Server
  - Microsoft SQL Server Integration Services
  - Pentaho Data Integration
- New challengers aiming at Big Data integration
  - Tamr Data Unification Platform

Setting Expectations

Alon Halevy: "Data Integration is AI-Complete"

• Meaning that completely automated solutions are unlikely.
• Reasons:
  1. System Level: Managing different platforms, distributed query processing
  2. Logical reasons: Schema and data heterogeneity
  3. Social reasons: Locating relevant data, convincing people to share (data fiefdoms)

Goal 1:
• Reduce the effort needed to set up an integration application.

Goal 2:
• Enable the system to perform gracefully with uncertainty (e.g., on the web)
Summary of Introduction

- Goal of Data Integration: Abstract away the fact that data comes from multiple sources in varying schemata
- The problem occurs everywhere: Handling it is curial for many applications in business, science, government, and Web
- Architectures range from warehousing to virtual integration
- Regardless of the architecture, bridging heterogeneity is the key issue
- Goal: Reduce the human effort involved
7. Course Outline and Organization
The Lecture

- introduces the principle methods of data integration
- discusses how to evaluate data integration results
- presents practical examples of how the methods are applied

- Topics
  1. Introduction to Data Integration
  2. Structured Data on the Web
  3. Data Exchange Formats
  4. Schema Mapping
  5. Identity Resolution
  6. Data Quality and Data Fusion
The Exercise

- Teams of **four students** realize a data integration project including
  1. data gathering
  2. schema mapping and data translation
  3. identity resolution
  4. data quality assessment and data fusion
- Teams will use data integration tools and will extend Java projects which implement basic integration methods
- Teams write 12 page report about their project, present project results
- You may choose their own application domain and data sets
  - minimum 4 data sets with a good degree of overlap in attributes and instances
- In addition, we will propose some suitable data sets from the domains of
  - films and actors, products and e-shops, restaurants, geographic information
- Overall grading
  - 50 % written final exam
  - 30 % written project report
  - 20 % presentation of project results
<table>
<thead>
<tr>
<th>Week</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.9.2015</td>
<td>Lecture: Introduction to Web Data Integration</td>
<td>Lecture: Structured Data on the Web</td>
</tr>
<tr>
<td>14.9.2015</td>
<td>Lecture: Data Exchange Formats</td>
<td>Lecture: Data Exchange Formats</td>
</tr>
<tr>
<td>5.10.2015</td>
<td>Feedback for Project Outlines</td>
<td>Exercise: Data Translation</td>
</tr>
<tr>
<td>12.10.2015</td>
<td>Exercise: Data Translation</td>
<td>Exercise: Data Translation</td>
</tr>
<tr>
<td>19.10.2015</td>
<td>Lecture: Identity Resolution</td>
<td>Lecture: Identity Resolution</td>
</tr>
<tr>
<td>26.10.2015</td>
<td>Exercise: Identity Resolution</td>
<td>Exercise: Identity Resolution</td>
</tr>
<tr>
<td>2.11.2015</td>
<td>Exercise: Identity Resolution</td>
<td>Exercise: Identity Resolution</td>
</tr>
<tr>
<td>9.11.2015</td>
<td>Lecture: Data Quality and Data Fusion</td>
<td>Lecture: Data Quality and Data Fusion</td>
</tr>
<tr>
<td>16.11.2015</td>
<td>Lecture: Data Quality and Data Fusion</td>
<td>Exercise: Data Fusion</td>
</tr>
<tr>
<td>23.11.2015</td>
<td>Exercise: Data Fusion</td>
<td>Exercise: Data Fusion</td>
</tr>
<tr>
<td>30.11.2015</td>
<td>Exercise: Data Fusion</td>
<td>Exercise: Data Fusion</td>
</tr>
<tr>
<td>7.12.2015</td>
<td>Presentation of Project Results</td>
<td>Presentation of Project Results</td>
</tr>
</tbody>
</table>
Course Organization

- Course Webpage
  - The lecture slides will be published on this webpage.
  - Project-related material will be provided in ILIAS.

- Time and Location
  - Wednesday, 15:30 to 17:00. Building: B6, Room: A 101
  - Thursday, 10:15 to 11:45. Building: B6, Room: A 101
  - Start: 10.9.2015
Waiting List

- The course is restricted to 30 participants.
- This semester we have **13 students on the waiting list**.
- If you are registered for the course
  - and decide not to attend anymore, please remove yourself from ILIAS so that somebody else can take your place.
  - please **sign into paper list today or tomorrow**, otherwise you will be removed next Wednesday
- If you are not registered for the course
  - Next week, might still assign you a place.
Literature and Credits


2. Ulf Leser, Felix Naumann: Informationsintegration. DBunkt Verlag, 2007. (Several copies in the library, also video lecture by Felix Naumann at http://www.tele-task.de/archive/series/overview/892/)

3. Xin Luna Dong, Divesh Srivastava: Big Data Integration, Morgan & Claypool, 2015 (Online access via the library)


Credits
The slide set of this lecture builds on slides from:
- Ulf Leser, Felix Naumann
- AnHai Doan, Alon Halevy, Zachary Ives
Lots of thanks to all of you!
Software

1. Data Translation
   - Altova MapForce
   - graphical data mapping and conversion tool

2. Identity Resolution
   - we will provide a Java project which implements the basic methods

3. Data Fusion
   - we will provide a Java project which implements the basic methods