Web Data Integration

Introduction to the Student Projects
Agenda

1. Overview
   - Phase I: Data Collection and Data Translation
   - Phase II: Identity Resolution
   - Phase III: Data Fusion

2. Details about Phase I: Data Collection and Data Translation
   - Requirements
   - Tool Support
   - Example

3. Group Formation

4. Start of Group Work
Overview Student Projects

• Phase I: Data Collection and Data Translation
  Duration: now – October 16th
  Tasks:
  1. Find a partner (groups of five)
  2. Decide on a use case
  3. Collect data from the Web
  4. Profile your data
  5. Generate integrated schema (target schema)
  6. Convert all your data into the integrated schema using MapForce

  Result: All data is represented using a single unified schema
  • one XML file per data source
Overview Student Projects

- **Phase II: Identity Resolution**
  
  Duration: October 17\textsuperscript{th} – November 12\textsuperscript{th}

  **Tasks:** Extend Java project template to
  
  1. Identify records in different data sets that describe the same entity
  2. Experiment with different combinations of similarity measures
  3. Use blocking to speed up the comparisons
  4. Evaluate quality of your approach

  **Result:** Correspondences between records in different data sets that describe the same entity
Overview Student Projects

• Phase III: Data Fusion
  Duration: November 13th – November 29th
  Tasks: Extend Java project template to
  1. Merge data and resolve data conflicts
  2. Experiment with different conflict resolution strategies
  3. Measure the quality and completeness of the final fused data set

Results:
  1. Fused data set in which each real-world entity is described by only a single record and these records contain no data conflicts
  2. Project report (12 pages) summarizing the results of the phases 1-3
Overview Student Projects

- Final Presentations
  - Dates: December 4th and December 5th
  - Overview of your use case
  - Explain your data
  - Explain the strategies that you used
  - Discuss the quality of your solution
Grading of the Projects (IE683, 3 ECTS)

Individual contribution to:

70%: Project work
- quality of your solution
- systematic experimentation with different alternatives
- systematic evaluation of experiments
- quality of written report

30%: Final presentation
- structure
- slides
- discussion

Please submit table on who did what together with the report.
Details about Phase I: Data Collection and Data Translation

• Duration: now – October 16\textsuperscript{th}

• Today
  1. Form teams of five people
  2. Decide on a domain/use case
  3. Start data collection and profiling

• Tomorrow
  1. Introduction to \textbf{MapForce}
  2. Start using MapForce to translate data to target schema

• Until Sunday, September 29\textsuperscript{th}, 23:59
  – Send a 4 page abstract on your project (details next slide)

• Wednesday, October 2\textsuperscript{nd}, 15:30-17:00
  – You get feedback on your abstract
Project Requirements

You should integrate:

1. at least **3 different data sets**
2. at least **2,500 entities** described in total (in joint dataset)
   - but more are better, good: >10,000 but <100,000
3. at least **1000 entities** should be contained in at least two datasets
   - please estimate based on small sample
4. at least **10 attributes** in joint dataset
   - entities should be identifiable by attribute combinations, e.g. name+birthdate
5. at least **5 attributes** should be contained in at least two datasets
   - some attributes should be contained in three datasets (for fusion by voting)
6. ideally, that datasets should contain a **list attribute**
   1. actors of a movie, directors of a company, songs on a CD
Project Abstracts

• Purpose of project abstract
  – check whether your ideas are feasible
  – proof that you fulfill the requirements (last slide)

• Content
  1. Brief description of use case
  2. Explanation how the datasets fulfill the requirements
    1. Schema and basic profile of each dataset
       • number of records per class
       • attributes with high percentage of missing values
    2. Integrated schema and overlap with input schemata
    3. Explanation why enough entities are likely contained in multiple datasets

• Submit via email to
  Anna Primpeli, Ralph Peeters and Christian Bizer

• Deadline: Sunday, September 29th, 23:59
Tables that MUST be used in Project Abstracts

1. Schema and Basic Profile of each Data Set

Table 1. Datasets

| Dataset   | Source (*) | format | class (**) | # of entities | # of attributes | list of attributes (***)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>Download URL</td>
<td>csv</td>
<td>movie</td>
<td>17,000</td>
<td>10</td>
<td>title, director, year, ...</td>
</tr>
<tr>
<td>DBpedia</td>
<td>dbpedia.org/sparql</td>
<td>xml</td>
<td>actor</td>
<td>23,500</td>
<td>8</td>
<td>name, birthDate (MV), activeYears,...</td>
</tr>
<tr>
<td>Freebase</td>
<td>Download URL</td>
<td>csv</td>
<td>actor</td>
<td>11,000</td>
<td>14</td>
<td>given_name, surname, spouse (MV)</td>
</tr>
</tbody>
</table>

(*) Should explain where from and how you got the data
(**) Add a line for each class, like in lines 1 and 2 of the example above
(***) Mark attributes with >30% missing values (MV)

2. Integrated Schema and Overlap with Input Schemata

Table 2. Attribute Intersection with Integrated Schema

<table>
<thead>
<tr>
<th>Class name</th>
<th>Attribute name</th>
<th>Datasets in which attribute is found</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie</td>
<td>name</td>
<td>dataset1, dataset2, dataset3, dataset4</td>
</tr>
<tr>
<td>movie</td>
<td>director</td>
<td>dataset1, dataset3</td>
</tr>
<tr>
<td>movie</td>
<td>year</td>
<td>dataset2, dataset3, dataset4</td>
</tr>
</tbody>
</table>

.....
Coaching Sessions

- Anna and Ralph will give you tips and answer questions concerning your project.
- Registration via email to Anna and Ralph is mandatory!
  - until Monday night!
  - including the questions that you like to discuss
- Anna and Ralph will assign you a time slot on Wednesday and inform you about the slot via email.
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.9.2019</td>
<td>Lecture: Data Exchange Formats</td>
<td>Lecture: Data Exchange Formats</td>
</tr>
<tr>
<td>25.9.2019</td>
<td>Project: Introduction to Student Projects</td>
<td>Tool Intro: MapForce</td>
</tr>
<tr>
<td>02.10.2019</td>
<td>Project: Feedback about Project Outlines</td>
<td>- Holiday -</td>
</tr>
<tr>
<td>09.10.2019</td>
<td>Project Work: Data Translation</td>
<td>Lecture: Identity Resolution</td>
</tr>
<tr>
<td>23.10.2019</td>
<td>Coaching: Identity Resolution</td>
<td>Project Work: Identity Resolution</td>
</tr>
<tr>
<td>30.10.2019</td>
<td>Coaching: Identity Resolution</td>
<td>- Holiday -</td>
</tr>
<tr>
<td>06.11.2019</td>
<td>Lecture: Data Fusion</td>
<td>Lecture: Data Fusion</td>
</tr>
<tr>
<td>08.11.2019</td>
<td>Excursion to SAP in Walldorf. Topic: Data Integration @ SAP</td>
<td></td>
</tr>
<tr>
<td>13.11.2019</td>
<td>Tool Intro: Winte.r Data Fusion</td>
<td>Project Work: Data Fusion</td>
</tr>
<tr>
<td>20.11.2019</td>
<td>Coaching: Data Fusion</td>
<td>Project Work: Data Fusion</td>
</tr>
<tr>
<td>27.11.2019</td>
<td>Coaching: Data Fusion</td>
<td>Project Work: Data Fusion</td>
</tr>
<tr>
<td>04.12.2019</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>12.12.2019</td>
<td>Final Exam</td>
<td></td>
</tr>
</tbody>
</table>
Possible Use Cases for Student Projects

- **Movies**
  - directors, budget, actors, oscar nominations...

- **Musicians**
  - first name, last name, birth date, birth place, bands, albums …

- **Companies**
  - performance, sector, key persons, panama papers data

- **Statistical Data**
  - countries, regions, cities, population, area, leader, GDP, …

- **Books**
  - title, author(s), number of pages, language, publisher, translator, …
Example Use Case 1: Movies

• Individual Data Sets contain
  – Movies
  – Actors
  – Directors
  – Oscar Nominations & Wins
  – Golden Globe Nominations & Wins

• Integrated dataset will contain
  1. Movies with release date, budget,... and awards nominated/won
  2. lists of actors and directors per movie
Example Use Case 1: Movies

• Lists of Oscar/Golden Globe nominees and winners
  – http://aggdata.com/awards/oscar
  – http://aggdata.com/awards/golden_globes

• List of The Guardian greatest films (by Genre)

• A large movie list
Example Use Case 1: Movies

- Movie data from DBpedia
- Issue a SPARQL query against http://dbpedia.org/sparql
- Result can be stored as CSV, JSON, XML, ...

```sql
SELECT ?title ?budget ?gross ?director
WHERE {
  ?x a dbo:Film .
  ?x dbo:director ?d .
  FILTER(LANG(?title)="en")
}
```
Example Use Case 2: Statistical Data

- Statistics about countries / regions / cities from CIA World Fact Book e.g., health data, transportation data, …

- European countries and regions:
  - https://ec.europa.eu/eurostat/data/database

- Country and city data from Geonames:
  - http://download.geonames.org/export/dump/countryInfo.txt
  - http://download.geonames.org/export/dump/cities15000.zip

- ...and of course, you can add data from DBpedia

- Note: there are only ~200 countries
  - thus, you'll have to add other entities to make it >2,500, e.g., cities
Where do I find Data for my Project?

- **Google Dataset Search**
  - [https://toolbox.google.com/datasetsearch](https://toolbox.google.com/datasetsearch)
Where do I find Data for my Project?

- **Google Table Search**
  - [http://research.google.com/tables](http://research.google.com/tables)
  - Press “Export to Fusion Tables”
  - Press “See table”
  - Use “File”->”Download” to download a CSV

- **HTML Tables in General**
  - e.g., Firefox plugin “Export CSV”
  - Exports a table to CSV on right click
Where do I find Data for my Project?

- Portal listing and monitoring 260 data catalogs
  - [http://data.wu.ac.at/portalwatch/](http://data.wu.ac.at/portalwatch/)
Where do I find Data for my Project?

- Web APIs
  - e.g., programmableweb.com – lists almost 17,000 APIs
  - requires some additional effort (using the API and getting the data)
Where do I find Data for my Project?

- **DBpedia and other Linked Data sources**
- Look at a single resource
- Look which properties are there (preferable dbpedia-owl)
- Construct a SPARQL query
- Go to [http://dbpedia.org/sparql](http://dbpedia.org/sparql) and get the data
- **Hint:** use **OPTIONAL** for properties that are not present for all entities:

  ```sparql
  SELECT ?title ?budget ?gross ?director
  WHERE {
    ?x a dbo:Film .
    OPTIONAL {?x dbo:gross ?gross .}
    ...
  }
  ```

There are 87,000 Films in DBpedia, but only 9,000 with gross
Where do I find Data for my Project?

- **Schema.org data** that has been crawled from multiple web sites.
  - Product, local business, hotel, job posting, ....
- [http://www.webdatacommons.org/structureddata/](http://www.webdatacommons.org/structureddata/)

### Class-Specific Subsets of the Schema.org Data

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Total Number of Quads:</th>
<th>Top Classes (Entity Count)</th>
<th>Total File Size</th>
<th>Quad File</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://schema.org/Airport">http://schema.org/Airport</a></td>
<td>80,258,863</td>
<td><a href="http://schema.org/Airport">http://schema.org/Airport</a> (28,764,415) <a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a> (9,238) <a href="http://schema.org/Product">http://schema.org/Product</a> (1,290) <a href="http://schema.org/Offer">http://schema.org/Offer</a> (1,783)</td>
<td>961 MB</td>
<td><a href="http://schema.org/Airport.nq.gz">sample</a></td>
</tr>
<tr>
<td><a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a></td>
<td>776,573,609</td>
<td><a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a> (48,086,763) <a href="http://schema.org/LocalBusiness">http://schema.org/LocalBusiness</a> (16,641,260) <a href="http://schema.org/GeoCoordinates">http://schema.org/GeoCoordinates</a> (12,345,942) <a href="http://schema.org/Place">http://schema.org/Place</a> (9,071,774)</td>
<td>14,354 MB</td>
<td><a href="http://schema.org/PostalAddress.nq.gz">sample</a></td>
</tr>
</tbody>
</table>
Creating an Integrated Schema

1. Have a look at your input data
   • Which entities exist? What attributes do they have?

2. Check input data against project requirements (see Slide 9)
   • Create the tables for the project abstract (see Slide 11)

3. Apply schema integration method from lecture (Spaccapietra, et al.)

E.g.
   – Movie: title, date, budget, revenue, oscar...
   – Actor/Director: first name, last name, birth date, nationality, ...

```
Movie       Actor (n:m)  Director(n:1)  Person
```
Hint: Create an example XML file
- using the integrated schema
- for some data from each input source
- in order to check if integrated schema can represent input data.

```xml
<movies>
    <movie>
        <title>2001</title>
        <director>
            <firstname>Stanley</firstname>
            <lastname>Kubrick</lastname>
        </director>
    </movie>
    ...
</movies>
```
Outlook: Exercise Tomorrow

1. Introduction to MapForce by Anna and Ralph
2. Start translating your data into the unifying schema using MapForce
Registration for SAP Analytics and Warehousing Day

- **Date:** Friday, November 8th, 9:30-17:00
- **Location:** SAP Headquarters Walldorf
- **Program:**

  09:30 – 10:00 Uhr: Check In, Ankunft (WDF01, 6. Stock, Raum Biel)
  10:00 – 11:00 Uhr: SAP Analytics & Data Management
  11:00 – 12:30 Uhr: SAP Data Warehousing & Data Management & Life at SAP
  12:30 – 13:30 Uhr: Gemeinsames Mittagessen in der SAP Foodhall
  13:30 – 15:00 Uhr: Example Showcase / Hands-on Exercise
  15:00 – 17:00 Uhr: Inspiration Pavilion “Augmented Intelligence” & Präsentation “SAP als Arbeitgeber”
...and now

1. Team formation
2. SAP Day Registration
3. Agree on use case
4. Start collecting data