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 17th 2018

  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 18\textsuperscript{th} – November 7\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
• **Phase III: Data Fusion**

  Duration: November 8\textsuperscript{th} – December 1\textsuperscript{st}

  **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 5th and December 6th
  - 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 17\textsuperscript{th} 2017

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

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

• **Until Sunday, October 7\textsuperscript{th}, 23:59**
  – Send a 4 page abstract on your project (details next slide)

• **Wednesday, October 10\textsuperscript{th}**
  – 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 **2 classes** in target schema
   - e.g., movie and actor
4. at least **1000 entities** should be contained in at least **two datasets**
   - please estimate based on small sample
5. at least **10 attributes** in joint dataset
   - entities should be identifiable by attribute combination, e.g. name+birthdate
6. at least **5 attributes** should be contained in at least **two datasets**
   - some attributes should be contained in three datasets (for fusion by voting)
7. at least **2 different input file formats**
   - CSV, JSON, XML...
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 fulfil 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, Oliver Lehmberg and Christian Bizer

• Deadline: Sunday, October 7th, 23:59
Tables for 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 (***)
|----------|------------|--------|------------|---------------|-----------------|------------------------|
| IMDB     | Download URL | csv    | movie      | 17,000        | 10             | title, director, year, ...
| DBpedia  | dbpedia.org/sparql | xml   | actor      | 23,500        | 8              | name, birthDate (MV), activeYears,...
| Freebase  | Download URL | csv    | actor      | 11,000        | 14             | given_name, surname, spouse (MV)

(*) 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>

.....
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.9.2018</td>
<td>Lecture: Introduction to Web Data Integration</td>
<td>Lecture: Structured Data on the Web</td>
</tr>
<tr>
<td>12.9.2018</td>
<td>Lecture: Data Exchange Formats</td>
<td>Lecture: Data Exchange Formats</td>
</tr>
<tr>
<td>26.9.2018</td>
<td>Project: Introduction to Student Projects</td>
<td>Tool Intro: MapForce</td>
</tr>
<tr>
<td>3.10.2018</td>
<td>- Holiday -</td>
<td>Project Work: Data Translation</td>
</tr>
<tr>
<td>10.10.2018</td>
<td>Project: Feedback about Project Outlines</td>
<td>Lecture: Identity Resolution</td>
</tr>
<tr>
<td>17.10.2018</td>
<td>Lecture: Identity Resolution</td>
<td>Tool Intro: Winte.r Identity Resolution</td>
</tr>
<tr>
<td>31.10.2018</td>
<td>Coaching: Identity Resolution</td>
<td>- Holiday -</td>
</tr>
<tr>
<td>7.11.2018</td>
<td>Lecture: Data Fusion</td>
<td>Lecture: Data Fusion</td>
</tr>
<tr>
<td>14.11.2018</td>
<td>Tool Intro: Winte.r Data Fusion</td>
<td>Project Work: Data Fusion</td>
</tr>
<tr>
<td>21.11.2018</td>
<td>Coaching: Data Fusion</td>
<td>Project Work: Data Fusion</td>
</tr>
<tr>
<td>28.11.2018</td>
<td>Coaching: Data Fusion</td>
<td>Project Work: Data Fusion</td>
</tr>
<tr>
<td>5.12.2018</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>17.12.2018</td>
<td>Final Exam</td>
<td></td>
</tr>
</tbody>
</table>
Coaching Sessions

- Anna and Oliver will give you tips and answer questions concerning your project.
- Registration via email to Anna and Oliver is mandatory!
  - until Monday night!
  - including the questions that you like to discuss
- Oliver will assign you a time slot on Wednesday and inform you about the slot via email.
Possible Use Cases for Student Projects

• Movies and Actors
  – Actors, directors, budget, oscar nominations...

• Musicians and Bands
  – First name, last name, birth date, birth place, bands, albums …

• Companies and Key Persons
  • company registries, panama papers data

• Statistical Data
  – Countries, regions, cities, population, area, leader, GDP, …

• Books and Authors
  – 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. Actors and directors with birth dates
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, …

```
SELECT ?title ?budget ?gross ?director
WHERE {
    ?x a dbpedia-owl:Film .
    ?x dbpedia-owl:gross ?gross .
    ?x dbpedia-owl: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:

• 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
Where do I find Data for my Project?

- **Google Table Search**
  - 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/
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:

```
SELECT ?title ?budget ?gross ?director
WHERE { ?x a dbpedia-owl:Film .
    OPTIONAL {?x dbpedia-owl: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</th>
<th>Top Classes (Entity Count)</th>
<th>Total File Size</th>
<th>Quad File</th>
</tr>
</thead>
</table>
| **http://schema.org/AdministrativeArea** | 1,724,857       | - http://schema.org/AdministrativeArea (100,671)  
- http://schema.org/GeoCoordinates (84,152)  
- http://schema.org/Country (63,851)  
- http://schema.org/Continent (83,567) | 23 MB           | [schema.orgAdministrativeArea.nq.gz (sample)]                                             |
| **http://schema.org/Airport**       | 80,268,863      | - http://schema.org/Airport (26,764,415)  
- http://schema.org/PostalAddress (9,238)  
- http://schema.org/Product (1,200)  
- http://schema.org/Offer (1,783)       | 961 MB          | [schema.orgAirport.nq.gz (sample)]                                                        |
| **http://schema.org/PostalAddress** | 776,573,609     | - http://schema.org/PostalAddress (48,086,763)  
- http://schema.org/LocalBusiness (16,641,260)  
- http://schema.org/GeoCoordinates (12,345,942)  
- http://schema.org/Place (9,071,774)      | 14,354 MB       | [schema.orgPostalAddress.nq.gz (sample)]                                                  |
| **http://schema.org/Product**       | 2,829,523,589   | - http://schema.org/Product (287,815,069)  
- http://schema.org/Offer (221,781,710)  
- http://schema.org/AggregateRating (38,398,548)  
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, ...

![Diagram showing relationships between Movie, Actor/Director, and Person]
Creating an Integrated Schema

**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

1. Introduction to MapForce
2. Translate your data into the unifying schema using MapForce
...and now

1. Find a partner
2. Agree on use case
3. Start collecting data