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 till October 27th

  Tasks:
  1. Find a partner (groups of four)
  2. Decide on a use case
  3. Collect data from the Web
  4. Profile your data and write outline about profile
  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 27\(^{th}\) – November 17\(^{th}\)
  - **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 17\(^{th}\) – December 9\(^{th}\)
  
  **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 8th and December 9th
  - Overview of your use case
  - Explain your data
  - Explain the strategies that you used in each step
  - Discuss the quality of your solution of each step
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.
Coaching Sessions

- Anna and Alexander will give you tips and answer questions concerning your project.

- Registration via email to Anna and Alexander is mandatory!
  - until Monday night (min. 2 days before the coaching session)!
  - including the questions that you like to discuss

- Anna and Alexander will assign you a time slot for the Wednesdays/Thursdays coaching session 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>08.09.2021</td>
<td>Lecture: Introduction to Web Data Integration</td>
<td>Lecture: Structured Data on the Web</td>
</tr>
<tr>
<td>15.09.2021</td>
<td>Lecture: Data Exchange Formats</td>
<td>Q&amp;A: Data Exchange Formats</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exercise: Data Exchange Formats</td>
</tr>
<tr>
<td>29.09.2021</td>
<td>Project: Introduction to Student Projects</td>
<td>Project Work: Preparation of Project Outlines</td>
</tr>
<tr>
<td>06.10.2020</td>
<td>Project: Feedback about Project Outlines</td>
<td>Exercise: Introduction to MapForce</td>
</tr>
<tr>
<td>20.10.2020</td>
<td>Lecture: Identity Resolution</td>
<td>Q&amp;A: Identity Resolution</td>
</tr>
<tr>
<td>27.10.2020</td>
<td>Exercise: Identity Resolution</td>
<td>Project Work: Identity Resolution</td>
</tr>
<tr>
<td>03.11.2020</td>
<td>Coaching: Identity Resolution</td>
<td>Project Work: Identity Resolution</td>
</tr>
<tr>
<td>10.11.2020</td>
<td>Lecture: Data Quality and Data Fusion</td>
<td>Lecture: Data Quality and Data Fusion</td>
</tr>
<tr>
<td>17.11.2020</td>
<td>Q&amp;A: Data Quality and Data Fusion</td>
<td>Exercise: Data Quality and Data Fusion</td>
</tr>
<tr>
<td>24.11.2020</td>
<td>Project Work: Data Quality and Fusion</td>
<td>Coaching: Data Quality and Fusion</td>
</tr>
<tr>
<td>01.12.2021</td>
<td>Project Work: Data Quality and Fusion</td>
<td>Coaching: Data Quality and Fusion</td>
</tr>
<tr>
<td>08.12.2020</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>14.12.2020</td>
<td>Final Exam</td>
<td></td>
</tr>
</tbody>
</table>
Details about Phase I: Data Collection and Data Translation

• Duration: now – October 27th

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

• Until Sunday, October 3rd, 23:59
  – Send a 4 page abstract on your project (details next slide)

• Wednesday, October 6th, 15:30-17:00
  – You get feedback on your abstract (if necessary)

• Thursday, October 7th, 10:15-11:45
  1. Introduction to MapForce
  2. Start using MapForce to translate data to target schema
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 **8 attributes** in joint dataset
   - entities should be identifiable by attribute combinations of at least **two attributes**, e.g. name+birthdate
5. at least **5 attributes** should be contained in at least **two datasets**
   - some attributes (other than name) should be contained in three datasets (for fusion by voting)
6. at least one of your attributes is a **list attribute**
   - 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, Alexander Brinkmann and Christian Bizer

- **Deadline: Sunday, October 3rd, 23:59**
Requirements for the Final Project Report

● 12 pages (sharp!) – counted without title page, table of content, literature list
  ○ Every extra page (including appendix pages) will reduce your mark by 0.33

● Due to Sunday, 5th December 2021, 23:59
  ○ Send by email to Chris, Alex and Anna

● You must use the DWS master thesis layout (without Chapters)
  https://www.uni-mannheim.de/dws/teaching/thesis-guidelines/

● Also submit
  1. your code and
  2. (a subset) of your data

● Please cite sources properly if you use any
  ○ Preferred citation style [Author, year]
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 (MV), 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, 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>first name, surname, spouse,…</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>Attribute type</th>
<th>Datasets in which the attribute is found</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie</td>
<td>name</td>
<td>string</td>
<td>dataset1, dataset2, dataset3, dataset4</td>
</tr>
<tr>
<td>movie</td>
<td>director</td>
<td>string/list</td>
<td>dataset1, dataset3</td>
</tr>
<tr>
<td>movie</td>
<td>year</td>
<td>date</td>
<td>dataset2, dataset3, dataset4</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
Possible Use Cases for Student Projects

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

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

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

- **Songs**
  - title, album, artist, releases, producer, composer…

- **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, …

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

- Individual Data Sets contain a subset of the following attributes:
  - Title of the song
  - Album
  - Artist name or list of artists
  - Producer
  - Composer
  - Releases: A list of the different releases of the same song. The release attribute has further nested attributes:
    - Year
    - Length
    - Special occasion (e.g. Live concert)
Example Use Case 2: Songs

- **MusicBrainz Database**
  - https://musicbrainz.org/doc/MusicBrainz_Database/Download

- **Spotify data set of 160K+ Tracks**
  - https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

- **Million Song Dataset with songs and release information (use the subset)**
  - http://millionsongdataset.com/pages/getting-dataset/#subset
Example Use Case 3: Companies

• Individual Data Sets contain a subset of the following attributes:
  - company name
  - website
  - founding date
  - headquarters country
  - headquarters city
  - industry
  - assets
  - revenue
  - founders
Example Use Case 3: Companies

- Forbes data set with top 2000 companies worldwide

- Open Data 500 Companies data set with 500 US located companies

- Kaggle data set with 7.1M companies
  - [https://www.kaggle.com/kaleab1/companies](https://www.kaggle.com/kaleab1/companies)

- Companies data from DBpedia

```sql
SELECT ?name ?ind_label ?equity ?income
WHERE { ?x a dbo:Company .
  ?x rdfs:label ?name .
  ?x dbo:industry ?industry .
  FILTER(LANG(?name)="en" && LANG(? ind_label)="en")
}
```
Where do I find Data for my Project?

- Google Dataset Search
  - https://toolbox.google.com/datasetsearch
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?

• Kaggle hosts numerous data sets and data competitions
  – https://www.kaggle.com/
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, Wikidata 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 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 {
  }
...}
```

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><img src="http://schema.org/Airport" alt="schema.org/Airport" /></td>
<td>80,258,803</td>
<td><a href="http://schema.org/Airport">http://schema.org/Airport</a> (26,704,415) <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/Thing">http://schema.org/Thing</a> (1,783)</td>
<td>961 MB</td>
<td><a href="http://schema.org/Airport.nq.gz">schemaorgAirport.nq.gz</a> (sample)</td>
</tr>
<tr>
<td><img src="http://schema.org/PostalAddress" alt="schema.org/PostalAddress" /></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> (9,671,774)</td>
<td>14,364 MB</td>
<td><a href="http://schema.org/PostalAddress.nq.gz">schemaorgPostalAddress.nq.gz</a> (sample)</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 11)**
   - Create the tables for the project abstract (see Slide 13)

3. **Apply schema integration method from lecture**
   - Rules of Thumb or 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: Exercise Next Week

1. Introduction to MapForce by Anna and Alexander
2. Start translating your data into the unifying schema using MapForce
...and now

1. Team formation
   a. Students with team:
      □ Start team call now and plan your project
   b. Students without team
      □ Let us know and we will assign you

Are there any open issues?
Are there any questions?

2. Agree on use case

3. Start collecting data