Web Data Integration: Introduction to Student Projects
Introduction to Student Projects

• Agenda
  – Overview all three exercises
  – Exercise 1
    • Requirements
    • Tool Support
  – Example settings
  – Group Phase
    • Find a group
    • Define a use case
Overview Student Projects

• Three project phases
• Phase I: now – October 22\textsuperscript{th}
  – Find a partner (groups of two)
  – Decide on a use case
  – Collect data from the web
  – Profile your data
  – Define a target schema
  – Convert all your data to the target schema
Overview Student Projects

• Three project phases
• Phase II: October 23\textsuperscript{th} – November 12\textsuperscript{th}
  – Find duplicates in your data
  – Test different measures
  – Combine measures
  – Evaluate quality of your approach
Overview Student Projects

• Three project phases
• Phase III: November 13\textsuperscript{st} – November 28\textsuperscript{th}
  – Merge data in final schema
  – Experiment with different merging strategies
  – Evaluate the quality of merging strategies
  – Create your final dataset
  – Report the quality of the final dataset

• By the end of the project: submit a \textbf{final report}
  – 12 pages, datasets description, results of the exercises 1-3
Overview Student Projects

• Final Presentations: December 4th – December 5th
  – Show your use case
  – Explain your data
  – Explain the strategies you used
  – Discuss the quality of your solution
Student Projects: Grading

- 30%: project work
  - Demand of your solution
  - Quality of your solution
  - Number of alternatives examined
  - Quality of evaluation (i.e., written report)
- 20%: final presentation
  - Structure
  - Slides
  - Explanation
Overview Student Projects

• Phase I: **Data Translation**, now – October 22\textsuperscript{th}
• Today
  – Find a partner
  – Decide on a domain/use case
  – Start data collection
  – Compile basic data profile
• Tomorrow
  – Introduction to **MapForce**
  – Start with Schema/Data Translation
• Until Tuesday, September 30\textsuperscript{th}
  – Send a 2-3 page abstract on your planned project
• Next Thursday, October 2\textsuperscript{th}
  – you get feedback on your abstract
Overview Student Projects

• Purpose of project abstract
  – check whether your ideas are feasible
  – give you additional hints

• Content
  – Brief use case description
  – Which datasets you use
    • Source
    • Schema
    • First profiling results (size, number of attributes)

• Submit via mail to volha@informatik.uni-mannheim.de
  – Tuesday, September 30th, 23:59 the latest
Overview Student Projects

• Until October 22th
  – Define a unifying schema
  – Translate your data with MapForce
  – Profile your data
Possible Use Cases

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

• Countries
  – Population, area, leader, GDP, literacy rate...

• People
  – First name, last name, birth date, birth place, profession, …

• Books
  – Title, author(s), number of pages, language, publisher, translator, …
Possible Use Cases (ctd.)

• Music albums
  – Artist, title, record label, date, tracks, awards, …

• Buildings
  – City, address, coordinates, architect, height, floors, finishing date, …

• Public Funding in the EU
  – Receiver, amount, funding agency, purpose, date, …
Use Case Requirements

- At least 4 input datasets
- At least 2,500 instances in total (in joint dataset)
- At least 2 classes in target schema
  - e.g., movie and actor
- At least 50% of the instances contained in at least two datasets
- At least 10 attributes in joint dataset
  - At least 50% of the attribute values contained in at least two datasets
- At least 2 different input file formats
  - CSV, JSON, XML...
Example Use Case: Movies

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

• Joint dataset will contain
  – Movies with release date, budget,... and awards nominated/won
  – Actors and directors with birth dates
Example Use Case: 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: Movies

- Popular movies from Movie DB
  - Requires registration for an API Key!
  - Results in JSON
Example Use Case: 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 dbpedia-owl:Film .
  ?x dbpedia-owl:gross ?gross .
  ?x dbpedia-owl:director ?d .
  FILTER(LANG(?title)="en")
}
```
Example Use Case: Geographic Data

- Several data tables about countries from World Fact Book (CIA)
  - e.g., health data, transportation data, ...

- European countries and regions:

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

- City data (>15k inhabitants) from Geonames:
  - http://download.geonames.org/export/dump/cities15000.zip
Example Use Case: Geographic Data

- [http://www.nationsonline.org](http://www.nationsonline.org)
  - Various data on population, cities, …
- ...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
Example Use Case: Person Data

• Specific: Nobel prizes
• Nobel Prize Winners as Linked Data:
  – http://data.nobelprize.org/directory/laureate

• Data on Nobel Prize Winners:
  – http://aggdata.com/awards/nobel_prize_winners

• Wikipedia tables
  – First two search results on Google tables

• ...and of course: DBpedia
Searching for Data Sources

- Web data catalogs
  - e.g., datahub.io – lists more than 10,000 data sets
Searching for Data Sources

• Getting data from Web APIs
  – e.g., programmableweb.com – lists almost 10,000 APIs
  – requires some additional effort (using the API and getting the data)
Searching for Data Sources

• Google Tables:
  – http://research.google.com/tables
  – Search for tables (e.g., “oscar winning movies”)
  – Press “Import data” → “Import to Fusion Tables”
  – Press “See table”
  – Use “File”->”Download” to download a CSV

• Web tables in General
  – e.g., Firefox plugin “Export CSV”
  – Exports a table to CSV on right click
Searching for Data Sources

• DBpedia is a useful source for almost all kinds of data
• 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
• Optional: recursion, include more properties

• 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
Creating a Unifying Schema

- Have a look at your input data
  - Which entities exist?
  - What attributes do they have?
  - Maximal expressivity vs. maximal integration
    - e.g., first name and last name would be more expressive
    - but you should evaluate whether the conversion is always possible

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

```plaintext
Movie  Actor (n:m)  Person
       Director(n:1)
```
Creating a Unifying Schema

• Hint: create an example XML for your use case

```xml
<movies>
  <movie>
    <title>2001</title>
    <director>
      <firstname>Stanley</firstname>
      <lastname>Kubrick</lastname>
      ...
    </director>
  </movie>
  ...
</movies>
```
Data Profiling

• Create a basic profile of your data
  – Number of entities
  – Catalog of attributes
    • Name
    • Data type (number, date, string, ...)
    • For numbers: min, max, mean
    • % complete (i.e., how many instances have the attribute?)
  – Estimated overlap of instances/number of duplicates
    • Take small samples of each dataset
    • Try to find them manually
      – e.g., search for person name from dataset 1 in dataset 2

• A preliminary version should be included into your project abstract
  – Datasets size, number of attributes, rough estimate of the overlap
Outlook

• Tomorrow:
  – Translating your data into the unifying schema
  – Using MapForce
...and now

- Find a partner
- Define a use case
- Start collecting data