

# 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
  - Examples
  - Data Sources
3. Group Formation
4. Start of Group Work

# Overview Student Projects

- Phase I: Data Collection and Data Translation

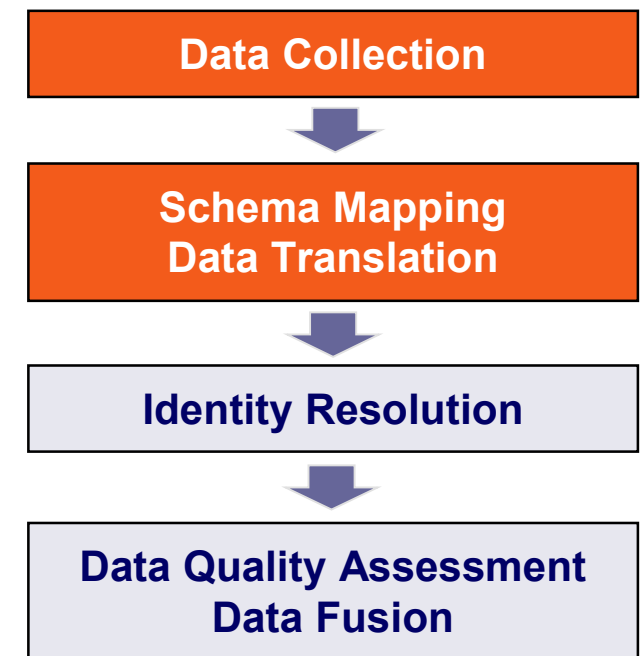
Duration: now till October 12<sup>th</sup>

**Tasks:**

1. Find a partner (groups of five)
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

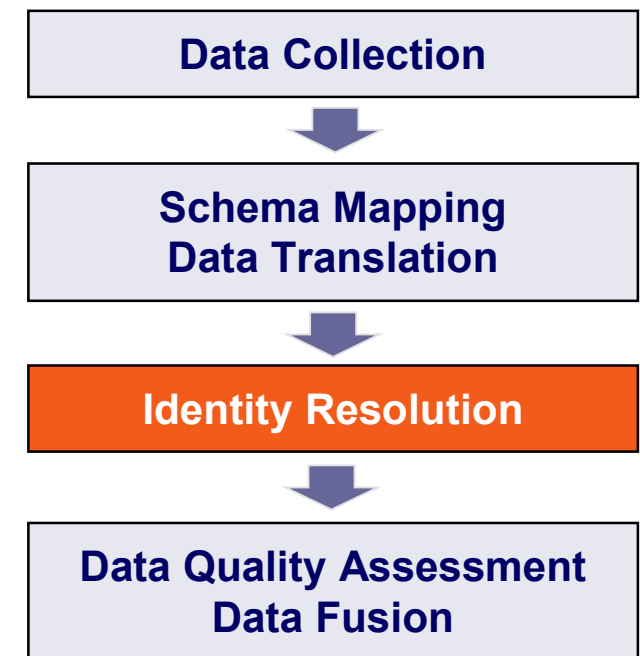
- **Phase II: Identity Resolution**

Duration: October 12<sup>th</sup> – November 9<sup>th</sup>

**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**

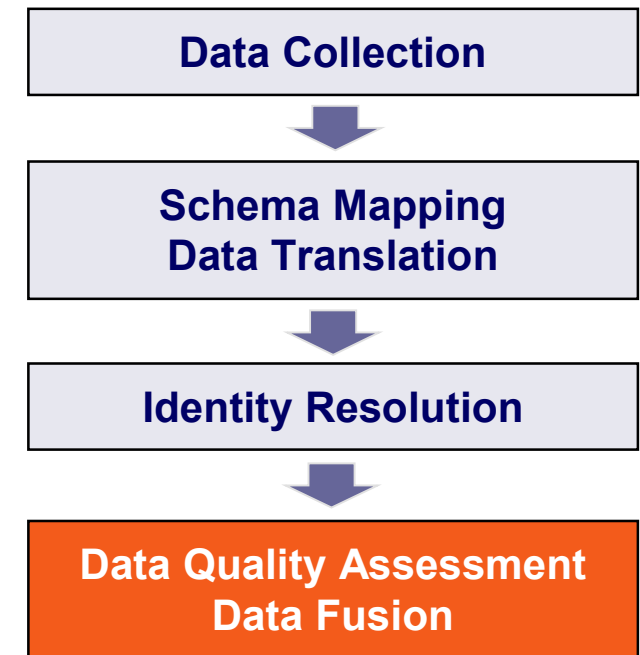
Duration: November 9<sup>th</sup> – December 1<sup>st</sup>

**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 7th and December 8th
- 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

- Ketī, Ralph and Alex will give you tips and answer questions concerning your project.
- Registration via email to Ketī, Ralph and Alex is mandatory!
  - until Monday night (min. 2 days before the coaching session)!
  - including the questions that you like to discuss
- Ketī, Ralph and Alex will assign you a time slot for the Wednesdays/Thursdays coaching session and inform you about the slot via email.



# Schedule

| Week       | Wednesday                                        | Thursday                                   |
|------------|--------------------------------------------------|--------------------------------------------|
| 07.9.2022  | Lecture: Introduction to Web Data Integration    | Lecture: Structured Data on the Web        |
| 14.9.2022  | Lecture: Data Exchange Formats                   | Lecture: Data Exchange Formats             |
| 21.9.2022  | Lecture: Schema Mapping                          | Lecture: Schema Mapping                    |
| 28.9.2022  | <b>Project: Introduction to Student Projects</b> | Exercise: Introduction to MapForce         |
| 05.10.2022 | Project: Feedback about Project Outlines         | Coaching: Schema Mapping                   |
| 12.10.2022 | Project Work: Schema Mapping                     | Lecture: Identity Resolution               |
| 19.10.2022 | Lecture: Identity Resolution                     | Exercise: Identity Resolution              |
| 26.10.2022 | Project Work: Identity Resolution                | Coaching: Identity Resolution              |
| 02.11.2022 | Project: Work Identity Resolution                | Coaching: Identity Resolution              |
| 09.11.2022 | Lecture: Data Quality and Data Fusion            | Lecture: Data Quality and Data Fusion      |
| 16.11.2022 | Exercise: Data Quality and Data Fusion           | Project Work: Data Quality and Data Fusion |
| 23.11.2022 | Project Work: Data Quality and Fusion            | Coaching: Data Quality and Fusion          |
| 30.11.2022 | Project Work: Data Quality and Fusion            | Coaching: Data Quality and Fusion          |
| 07.12.2022 | Presentation of Project Results (IE683)          | Presentation of Project Results (IE683)    |
| 15.12.2022 | Final Exam (IE670)                               |                                            |

# Details about Phase I: Data Collection and Data Translation

- Duration: now – October 12<sup>th</sup>
- **Today**
  1. Form teams of **five** people
  2. Decide on a domain/use case
  3. Start data collection and profiling
- **Until Monday, October 3<sup>rd</sup>, 23:59**
  - Send a 4 page abstract on your project (details next slide)
- **Wednesday, October 5<sup>th</sup>, 15:30-17:00**
  - You get feedback on your abstract (if necessary)
- **Thursday, September 29<sup>th</sup>, 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 **1,000 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. ideally, 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  
Keti Korini, Ralph Peeters, Alexander Brinkmann and Christian Bizer
- **Deadline: Monday, October 3<sup>rd</sup>, 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 (***)         |
|----------|--------------------|--------|-----------|---------------|-----------------|----------------------------------|
| IMDB     | Download URL       | csv    | Movie     | 17,000        | 10              | title, director (MV), year,...   |
| DBpedia  | Dbpedia.org/sparql | xml    | Actor     | 23,500        | 8               | name, birthDate, activeYears,... |
| Freebase | Download URL       | csv    | Actor     | 11,000        | 14              | first name, surname, spouse,...  |

(\*) 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

| Class name | Attribute name | Attribute type | Datasets in which the attribute is found                          |
|------------|----------------|----------------|-------------------------------------------------------------------|
| movie      | name           | string         | dataset1, dataset2, dataset3, dataset4 (use proper dataset names) |
| movie      | director       | string/ list   | dataset1, dataset3                                                |
| movie      | year           | date           | dataset2, dataset3, dataset4                                      |
| ...        | ...            |                | ...                                                               |

# 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, 4th December 2022, 23:59**
  - Send by email to **Chris, Ket, Ralph and Alex**
- You must use latex template of **Springer CS Proceedings**
  - <http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines>
- Also **submit**
  - your **code** and
  - (a subset) of your **data**
- Please cite sources properly if you use any
  - Preferred citation style [Author, year]

# 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](http://aggdata.com/awards/golden_globes)
  - <http://www.amstat.org/publications/jse/datasets/oscars.dat.txt>
- List of The Guardian greatest films (by Genre)
  - <http://www.guardian.co.uk/news/datablog/2010/oct/16/greatest-films-of-all-time>
- A large movie list
  - [https://github.com/vlandham/vlandham.github.com/blob/master/vis/movie/data/movies\\_all.csv](https://github.com/vlandham/vlandham.github.com/blob/master/vis/movie/data/movies_all.csv)



# 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 dbo:Film .
        ?x dbo:budget ?budget .
        ?x dbo:gross ?gross .
    ?x dbo:director ?d .
        ?d foaf:name ?director .
        ?x rdfs:label ?title .
    FILTER(LANG(?title)="en")
}
```

# Example Use Case 2: Companies

- Goal: Combine multiple datasets into a single dataset having the following attributes:
  - company name
  - website
  - founding date
  - headquarters country
  - headquarters city
  - industry
  - assets
  - revenue
  - founders

# Sources of Data about Companies

- Forbes data set with top 2000 companies worldwide
  - <https://www.kaggle.com/ash316/forbes-top-2000-companies>
- Open Data 500 Companies data set with 500 US located companies
  - <https://www.kaggle.com/govlab/open-data-500-companies>
- Kaggle data set with 7.1M companies
  - <https://www.kaggle.com/kaleab1/companies>
- Companies data from DBpedia

```
SELECT ?name ?ind_label ?equity ?income
WHERE {
  ?x a dbo:Company .
      ?x rdfs:label ?name .
      ?x dbo:industry ?industry .
      ?industry rdfs:label ?ind_label .
      ?x dbo:equity ?equity .
      ?x dbo:netIncome ?income .
                                     FILTER (LANG (?name) = "en"
                                     && LANG (? ind_label) = "en")
}
```

# Where do I find Data for my Project?

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



The image displays two screenshots of the Google Dataset Search interface. The left screenshot shows search results for the query "Company listing". The right screenshot shows search results for the query "Cities".

**Left Screenshot (Search: Company listing):**

- Business Registration Data** (data.ct.gov, Updated Nov 20, 2015)
- Italy IT: No of Listed Domestic Companies: Total** (www.ceicdata.com)
- EDGAR Database of SEC Filings** (data.wu.ac.at)

**Right Screenshot (Search: Cities):**

- Cities and Towns NTAD** (data.world, Updated Sep 11, 2018)
- World Cities** (hub.arcgis.com, Published Jun 30, 2013)
- World Cities** (www.kaggle.com, Updated Jun 13, 2017)

# Where do I find Data for my Project?

- Portal listing and monitoring 260 data catalogs
  - <http://data.wu.ac.at/portalwatch/>

**Open Data Portal Watch**  
Quality Assessment and Monitoring of 260 Open Data Portals

WU  
WIRTSCHAFTS  
UNIVERSITÄT  
WIEN VIENNA  
UNIVERSITY OF  
ECONOMICS  
AND BUSINESS

ODPW System Portals API Quality Measures License Search SPARQL About

260 Portals  
Filter ... (by URI, Software, ISO)

Portals List  
Stats  
Quality  
Dynamicity

| Portal                | API Homepage | Software     | Snapshots | Last Update                | Datasets | Resources |
|-----------------------|--------------|--------------|-----------|----------------------------|----------|-----------|
| data.edostate.gov.ng  | API Homepage | CKAN         | 64        | Last Sep 18 - Sep 24, 2017 | NONE     | 0         |
| opendata.brussels.be  | API Homepage | OpenDataSoft | 65        | Last Sep 18 - Sep 24, 2017 | NONE     | 0         |
| data.ohouston.org     | API Homepage | CKAN         | 65        | Last Sep 18 - Sep 24, 2017 | NONE     | 0         |
| data.cityofdeleon.org | API Homepage | Socrata      | 66        | Last Sep 18 - Sep 24, 2017 | 12       | 0         |

# 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)

API Search: movies [Get Latest APIs](#)

All [Advertising \(218\)](#) [Answers \(32\)](#) [Auctions \(3\)](#) [Backend \(106\)](#) [Blog Search \(11\)](#) [Blogging \(70\)](#) [Bookmarks \(42\)](#) [Calendar \(33\)](#) [Catalog \(5\)](#) [Chat \(53\)](#) [Database \(99\)](#) [Dating \(3\)](#) [Dictionary \(23\)](#) [Education \(199\)](#) [Email \(200\)](#) [Enterprise \(438\)](#) [Entertainment \(80\)](#) [Events \(99\)](#) [Fax \(22\)](#)

Filter this list ☒

See also: [Mashups tagged movies](#)

Sort **Name** Date Popularity Category

1 to 20 of 59 APIs

| API                           | Description                                      | Category      | Date       |
|-------------------------------|--------------------------------------------------|---------------|------------|
| 90sec.tv                      | Embed videos on web sites                        | Video         | 2010-11-01 |
| Alibris                       | Marketplace for books, movies and music          | Shopping      | 2010-12-03 |
| Allociné Movie                | French-language movie and TV information service | Entertainment | 2012-01-05 |
| AllSubs                       | Movie and television subtitle database           | Database      | 2013-09-09 |
| Arts Alliance Media Locksmith | Digital cinema KDM delivery service              | Security      | 2012-10-08 |
| Bol.com                       | Dutch online retail store                        | Shopping      | 2012-01-13 |
| CinemaTorrents                | Movie Torrent Search                             | Search        | 2011-01-18 |
| Cineworld                     | UK cinema information                            | Entertainment | 2013-05-07 |



# Where do I find Data for my Project?

- DBpedia, Wikidata and other Linked Data sources
- Look at a single resource
  - [http://dbpedia.org/resource/2001:\\_A\\_Space\\_Odyssey\\_%28film%29](http://dbpedia.org/resource/2001:_A_Space_Odyssey_%28film%29)
- 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:

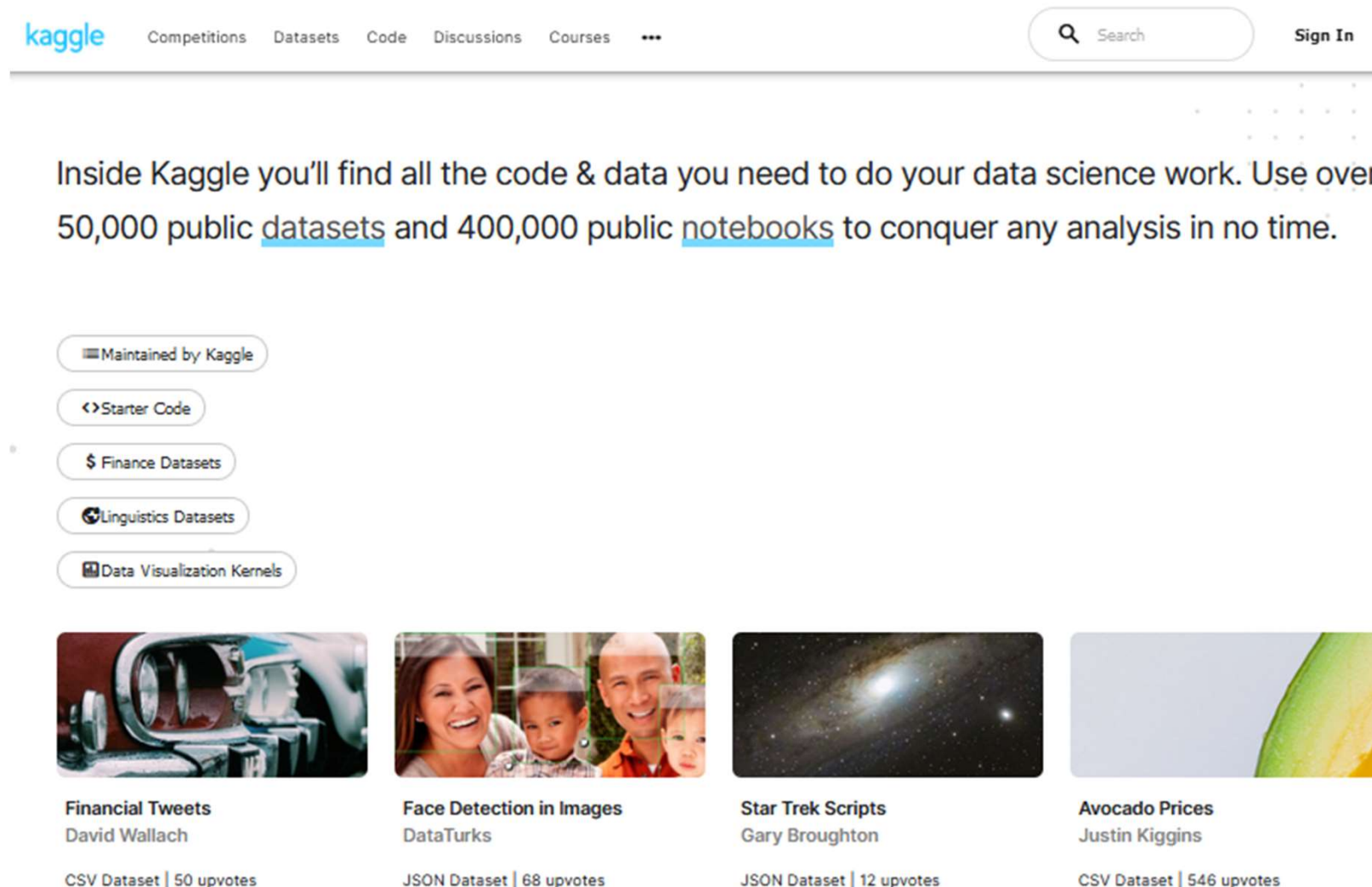
```
SELECT ?title ?budget ?gross ?director
WHERE {
    ?x a dbo:Film .
    OPTIONAL {?x dbo:gross ?gross . }
    ...
}
```

There are 87,000 Films  
in DBpedia, but only  
9,000 with gross



# Where do I find Data for my Project?

- Kaggle hosts numerous data sets and data competitions
  - <https://www.kaggle.com/>



# 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://webdatacommons.org/structureddata/schemaorgtables/>



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

| Class Name                                                                              | Total Number of                                            | Top Classes (Entity Count)                                                                                                                                                                                                                                                                                                                                            | Total File Size | Quad File                                                  |
|-----------------------------------------------------------------------------------------|------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|------------------------------------------------------------|
| <a href="http://schema.org/AdministrativeArea">http://schema.org/AdministrativeArea</a> | Quads: 1,724,857<br>URLs: 85,625<br>Hosts: 63              | <a href="http://schema.org/AdministrativeArea">http://schema.org/AdministrativeArea</a> (100,671)<br><a href="http://schema.org/GeoCoordinates">http://schema.org/GeoCoordinates</a> (84,152)<br><a href="http://schema.org/Country">http://schema.org/Country</a> (83,851)<br><a href="http://schema.org/Continent">http://schema.org/Continent</a> (83,567)         | 23 MB           | <a href="#">schemaorgAdministrativeArea.nq.gz (sample)</a> |
| <a href="http://schema.org/Airport">http://schema.org/Airport</a>                       | Quads: 80,258,863<br>URLs: 963,538<br>Hosts: 99            | <a href="http://schema.org/Airport">http://schema.org/Airport</a> (26,764,415)<br><a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a> (9,238)<br><a href="http://schema.org/Product">http://schema.org/Product</a> (1,290)<br><a href="http://schema.org/Offer">http://schema.org/Offer</a> (1,283)                                         | 961 MB          | <a href="#">schemaorgAirport.nq.gz (sample)</a>            |
| <a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a>           | Quads: 776,573,609<br>URLs: 13,475,055<br>Hosts: 131,064   | <a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a> (48,086,763)<br><a href="http://schema.org/LocalBusiness">http://schema.org/LocalBusiness</a> (16,641,260)<br><a href="http://schema.org/GeoCoordinates">http://schema.org/GeoCoordinates</a> (12,345,942)<br><a href="http://schema.org/Place">http://schema.org/Place</a> (9,071,774) | 14,364 MB       | <a href="#">schemaorgPostalAddress.nq.gz (sample)</a>      |
| <a href="http://schema.org/Product">http://schema.org/Product</a>                       | Quads: 2,829,523,589<br>URLs: 48,314,143<br>Hosts: 104,118 | <a href="http://schema.org/Product">http://schema.org/Product</a> (287,815,069)<br><a href="http://schema.org/Offer">http://schema.org/Offer</a> (221,781,710)<br><a href="http://schema.org/AggregateRating">http://schema.org/AggregateRating</a> (38,398,548)<br><a href="http://schema.org/Review">http://schema.org/Review</a> (26,209,678)                      | 62,179 MB       | <a href="#">schemaorgProduct.nq.gz (sample)</a>            |

# 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, ...

# Creating an Integrated Schema

**Hint:** Create an example XML file

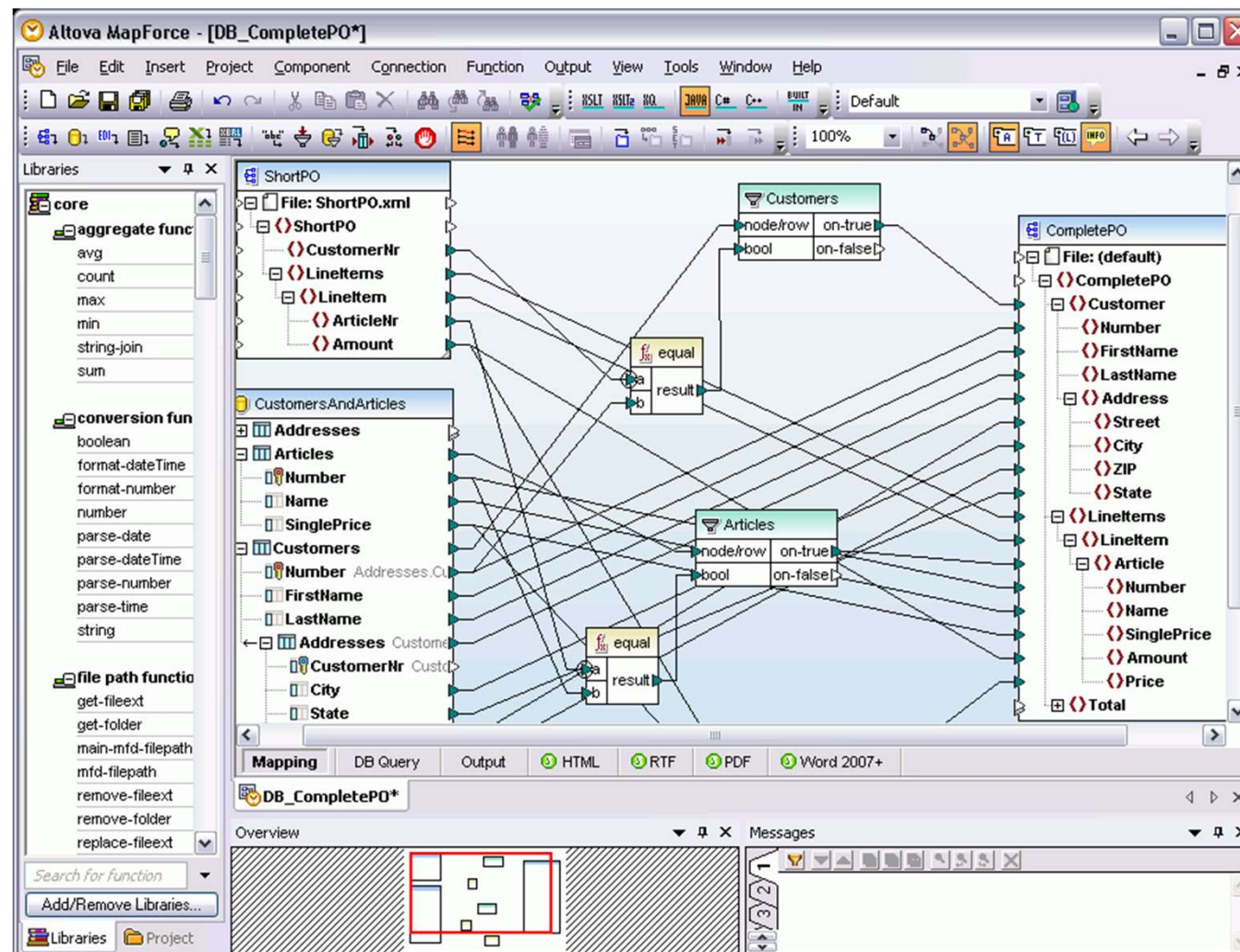
- using the integrated schema
- for some data from each input source
- in order to check if integrated schema can represent input data.

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



# Outlook: Exercise Next Week

1. Introduction to MapForce by Ketil and Ralph
2. Start translating your data into the unifying schema using MapForce



# ...and now

1. Team formation
  - a. Students with team:
    - ➔ Put your name in the list
  - 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

