

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

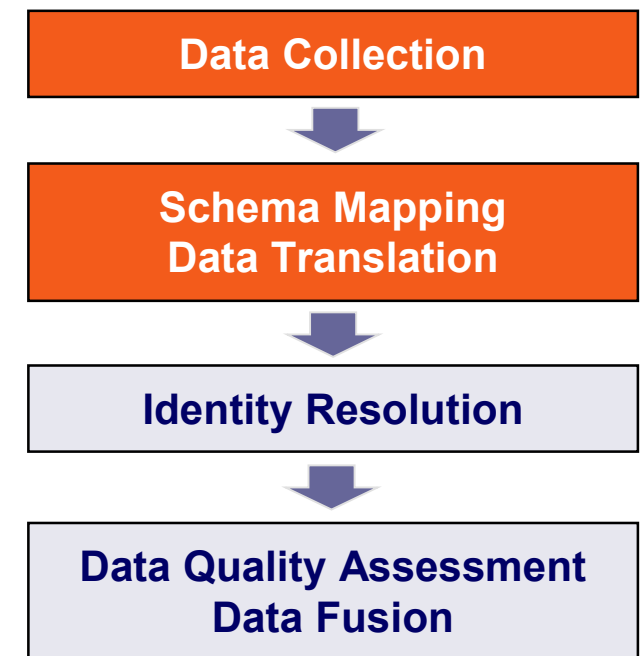
Duration: now till October 18th

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

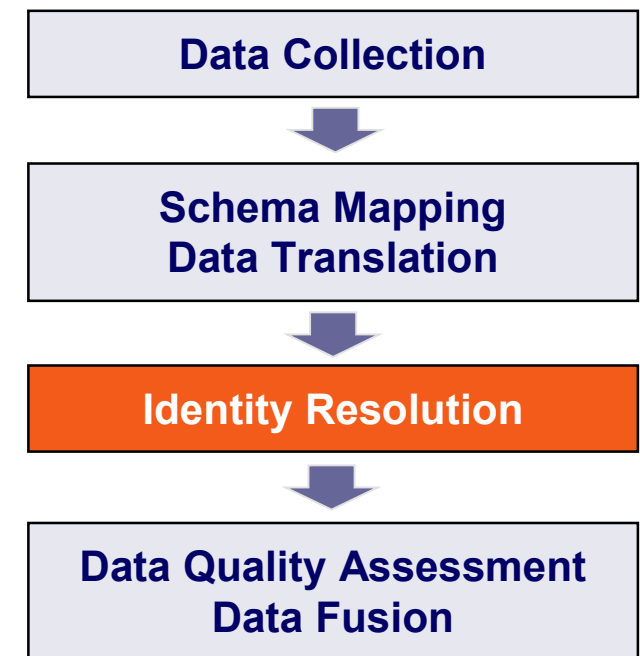
- **Phase II: Identity Resolution**

Duration: October 18th – November 8th

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**

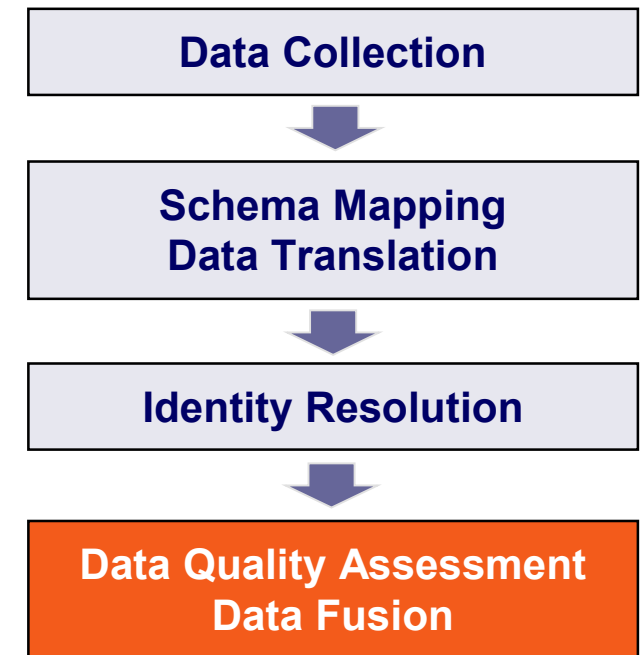
Duration: November 8th – November 30th

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 6th and December 7th
- 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
06.09.2023	- no lecture –	Lecture: Introduction to Web Data Integration
13.9.2023	Lecture: Structured Data on the Web	Lecture: Data Exchange Formats
20.9.2023	Lecture: Data Exchange Formats	Lecture: Schema Mapping
27.9.2023	Lecture: Schema Mapping	Project: Introduction to Student Projects
04.10.2023	Exercise: Introduction to MapForce	Coaching: Schema Mapping
11.10.2023	Project: Feedback about Project Outlines	Lecture: Identity Resolution
18.10.2023	Lecture: Identity Resolution	Exercise: Identity Resolution
25.10.2023	Project Work: Identity Resolution	Coaching: Identity Resolution
02.11.2023	- Public Holiday –	Coaching: Identity Resolution
08.11.2023	Lecture: Data Quality and Data Fusion	Lecture: Data Quality and Data Fusion
15.11.2023	Exercise: Data Quality and Data Fusion	Project Work: Data Quality and Data Fusion
22.11.2023	Project Work: Data Quality and Fusion	Coaching: Data Quality and Fusion
30.11.2023	Project Work: Data Quality and Fusion	Coaching: Data Quality and Fusion
06.12.2023	Presentation of Project Results (IE683)	Presentation of Project Results (IE683)
18.12.2023	Final Exam (IE670)	

Details about Phase I: Data Collection and Data Translation

- Duration: now – October 18th
- **Today**
 1. Form teams of **five** people
 2. Decide on a domain/use case
 3. Start data collection and profiling
- **Until Monday, October 9th, 23:59**
 - Send a 4 page abstract on your project (details next slide)
- **Wednesday, October 11th, 15:30-17:00**
 - You get feedback on your abstract (if necessary)
- **Wednesday, October 4th, 15:30-17:00**
 1. Introduction to **MapForce**
 2. Start using MapForce to translate data to target schema



Project Requirements

You should integrate:

1. 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 9th, 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	# of entities	# of attributes	List of attributes (**)
IMDB	Download URL	csv	17,000	10	title, director (MV), year,...
DBpedia	Dbpedia.org/sparql	xml	23,500	8	name, birthDate, activeYears,...
Freebase	Download URL	csv	11,000	14	first name, surname, spouse,...

(*) Should explain where from and how you got the data

(**) Mark attributes with >30% missing values (MV)

2. Integrated Schema and Overlap with Input Schemata

Table 2. Attribute Intersection with Integrated Schema

Attribute name	Attribute type	Datasets in which the attribute is found
name	string	dataset1, dataset2, dataset3, dataset4 (use proper dataset names)
director	string/list	dataset1, dataset3
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, 3rd December 2023, 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...
- Companies
 - performance, sector, key persons, Panama papers data
- Musicians
 - first name, last name, birth date, birth place, bands, albums ...
- 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://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



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 (regional branch vs. company)
 - headquarters city
 - industry (single taxonomy)
 - Assets (normalized)
 - revenue (normalized)
 - founders (list)

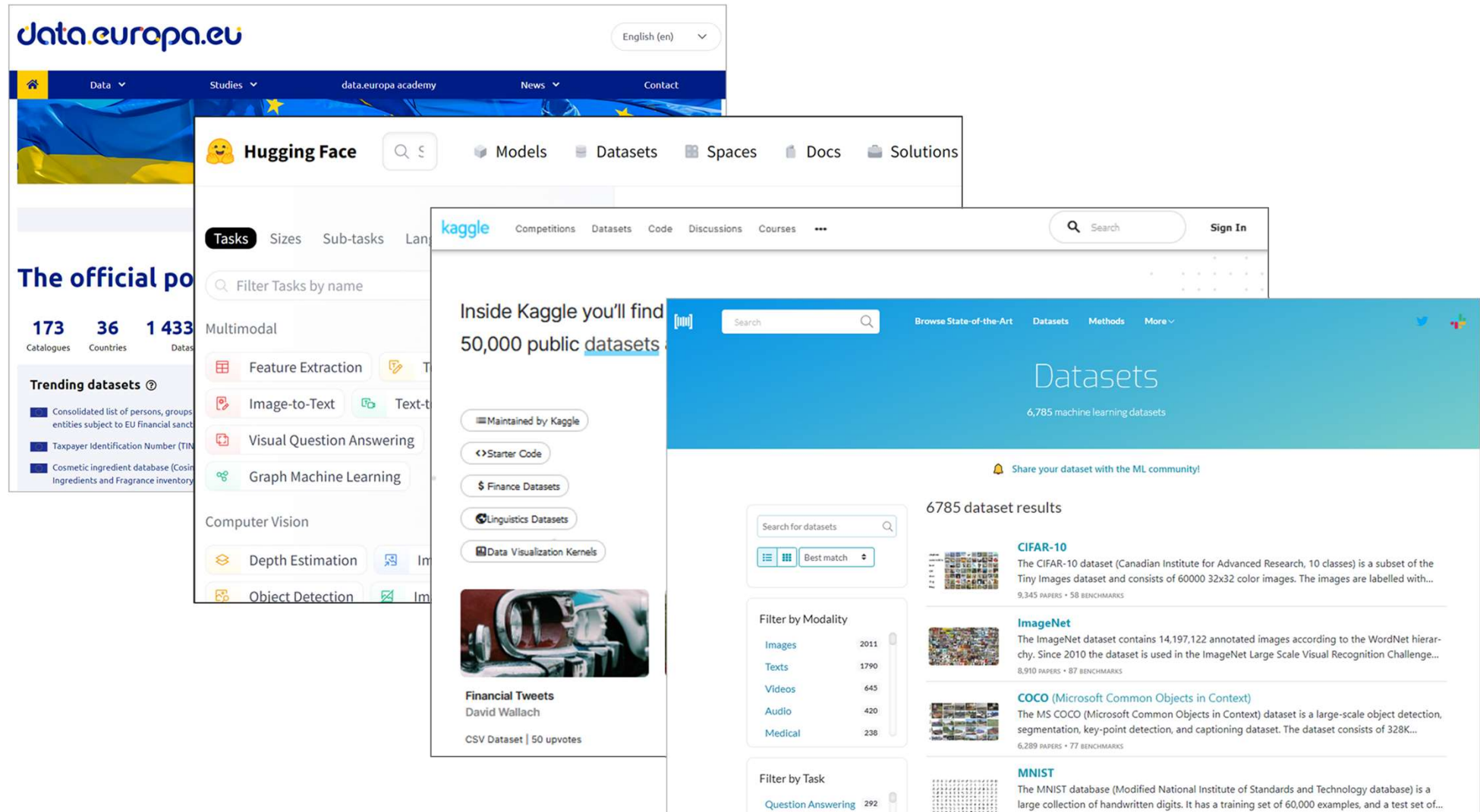
Sources of Data about Companies

- 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?

- Data Portals



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, illustrating how to find data for a project.

Left Screenshot: Search for 'Company listing'

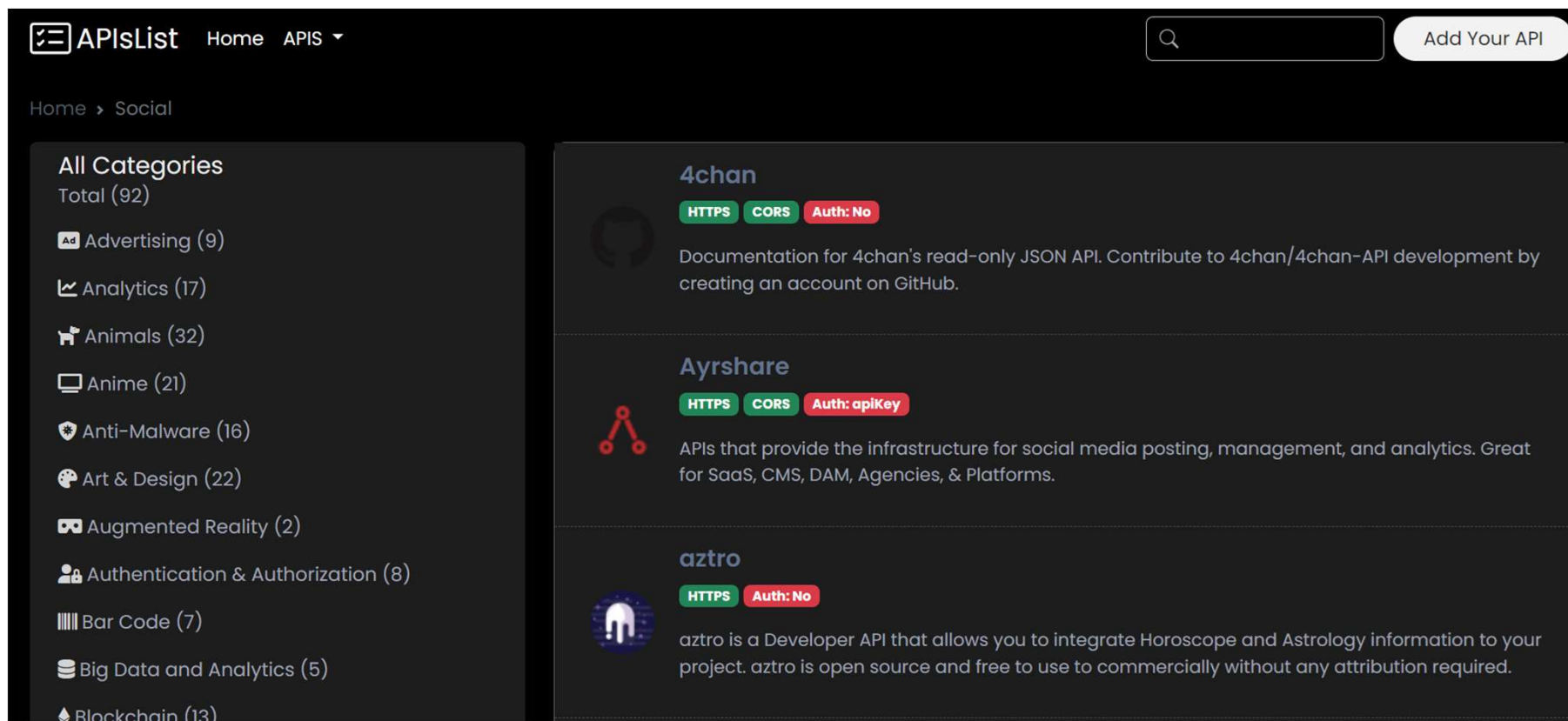
- Search Bar:** Google Dataset Search, Company listing
- Results:**
 - Business Registration Data** (data.ct.gov, Updated Nov 20, 2015)
 - Italy IT: No of Listed Domestic Companies: Total** (www.ceicdata.com)
 - EDGAR Database of SEC Filings** (data.wu.ac.at)

Right Screenshot: Search for 'Cities'

- Search Bar:** Google Dataset Search, Cities
- Results:**
 - 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?

- **Web APIs**
 - e.g., apislist.com – lists many public 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
 - 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?

- **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
http://schema.org/AdministrativeArea	Quads: 1,724,857 URLs: 85,625 Hosts: 63	http://schema.org/AdministrativeArea (100,671) http://schema.org/GeoCoordinates (84,152) http://schema.org/Country (83,851) http://schema.org/Continent (83,567)	23 MB	schemaorgAdministrativeArea.nq.gz (sample)
http://schema.org/Airport	Quads: 80,258,863 URLs: 963,538 Hosts: 99	http://schema.org/Airport (26,764,415) http://schema.org/PostalAddress (9,238) http://schema.org/Product (1,290) http://schema.org/Offer (1,283)	961 MB	schemaorgAirport.nq.gz (sample)
http://schema.org/PostalAddress	Quads: 776,573,609 URLs: 13,475,055 Hosts: 131,064	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)	14,364 MB	schemaorgPostalAddress.nq.gz (sample)
http://schema.org/Product	Quads: 2,829,523,589 URLs: 48,314,143 Hosts: 104,118	http://schema.org/Product (287,815,069) http://schema.org/Offer (221,781,710) http://schema.org/AggregateRating (38,398,548) http://schema.org/Review (26,209,678)	62,179 MB	schemaorgProduct.nq.gz (sample)

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, director, actors (list)

Creating an Integrated Schema

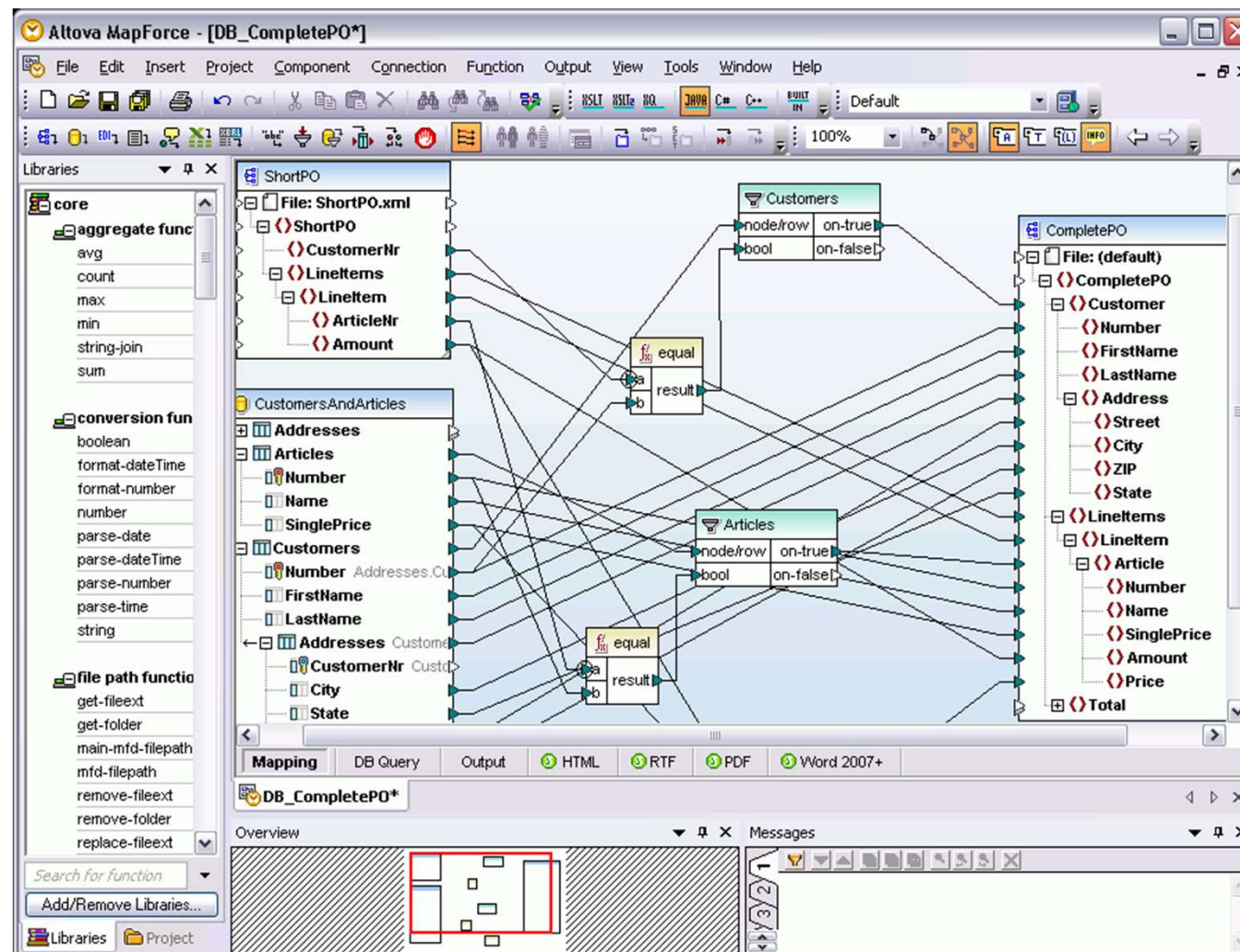
Hint: Create an example XML file

- 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 Wednesday

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

