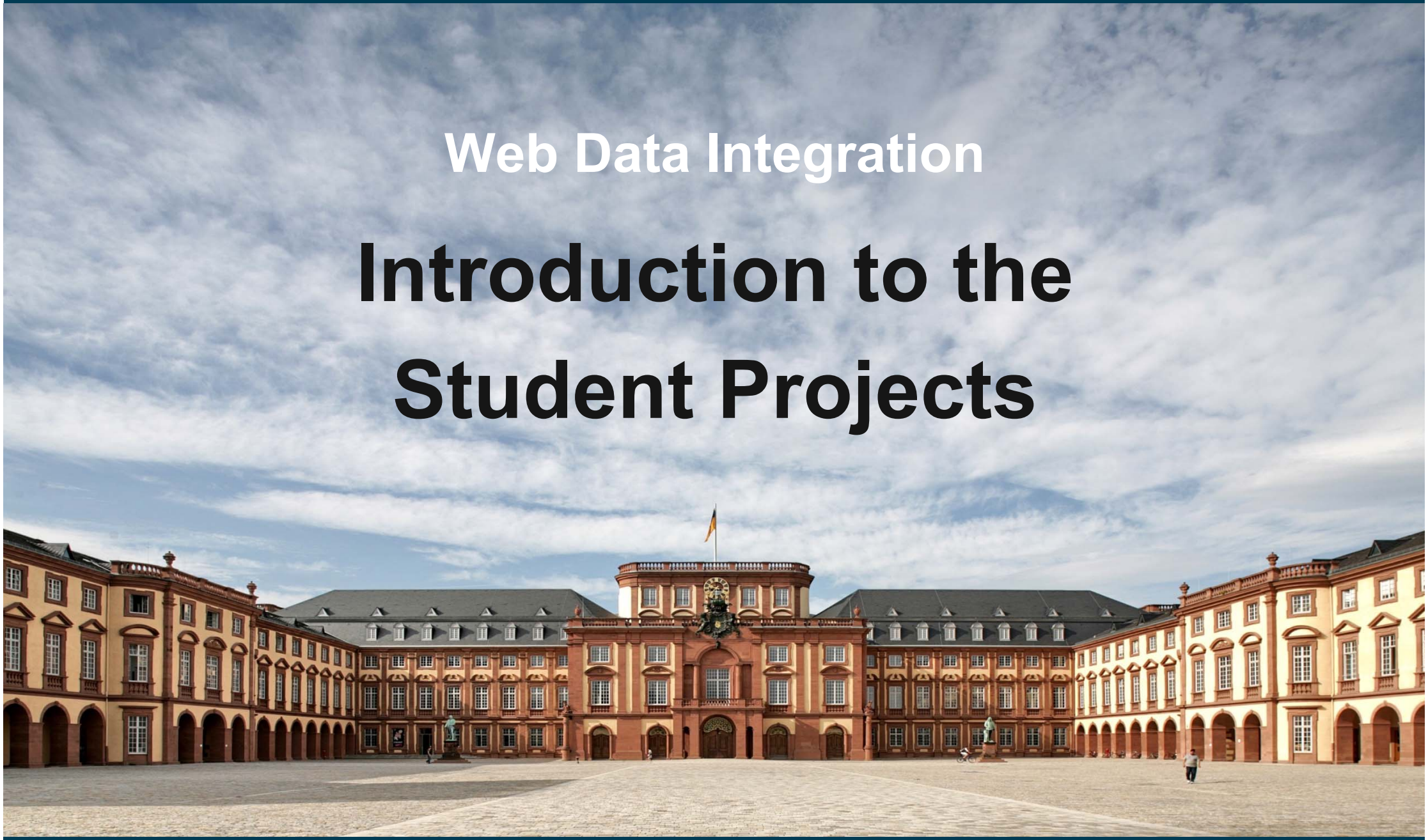
UNIVERSITY OF MANNHEIM

Web Data Integration

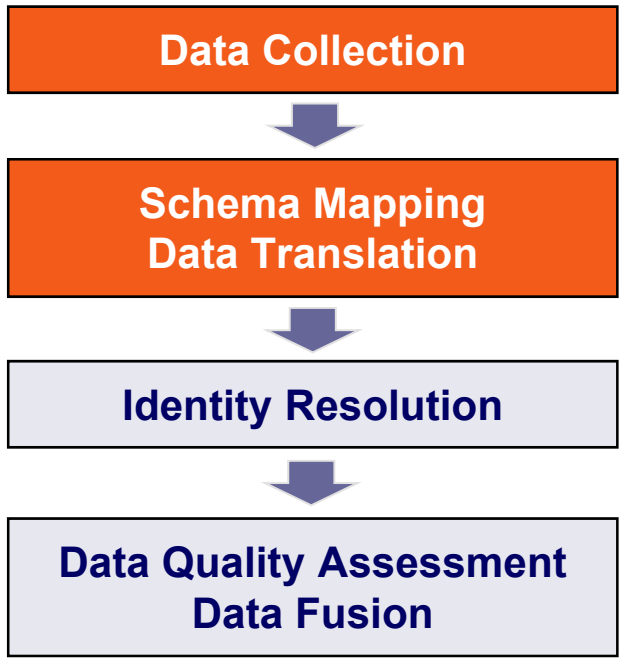
Introduction to the Student Projects



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 November 11th

  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

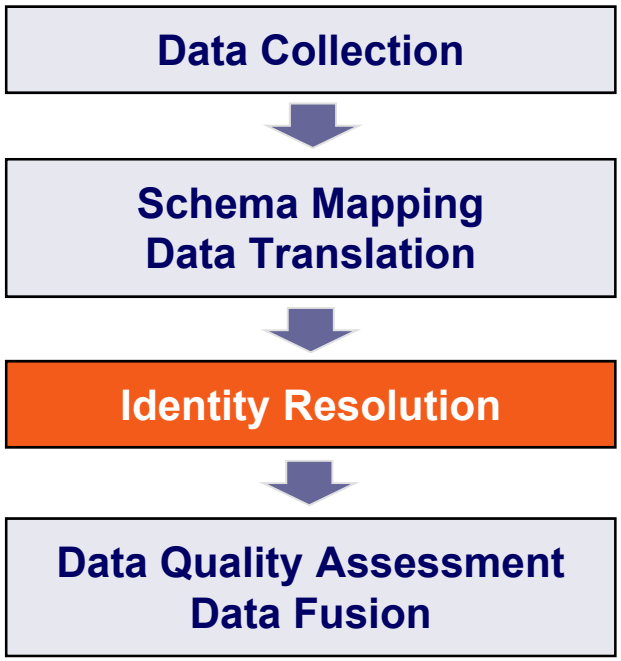
- Phase II: Identity Resolution

  Duration: November 11th – November 25th

  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



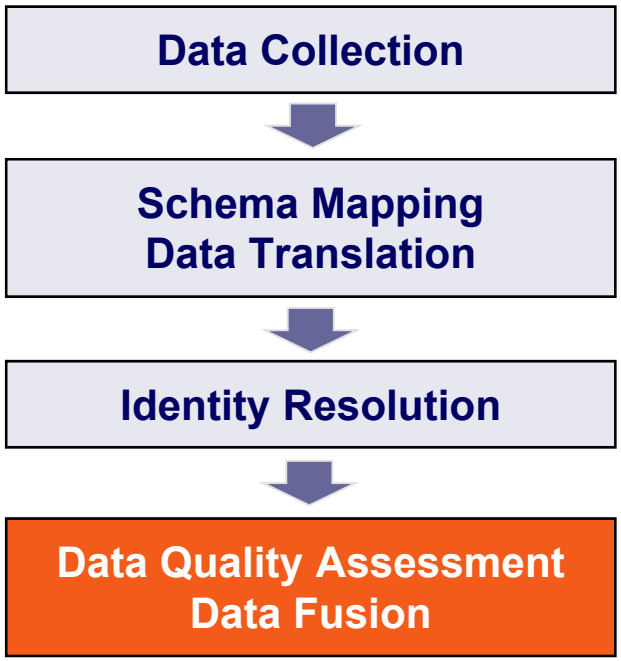
Overview Student Projects

- Phase III: Data Fusion

  Duration: November 25th – December 9th

  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 9th and December 10th
  - 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 coaching session Wednesdays and inform you about the slot via email.

Schedule

Week	Wednesday	Thursday
30.9.2020	Lecture: Introduction to Web Data Integration	Lecture: Structured Data on the Web
07.10.2020	Lecture: Data Exchange Formats	Q&A: Data Exchange Formats Exercise: Data Exchange Formats
14.10.2020	Lecture: Schema Mapping	Q&A: Schema Mapping
21.10.2020	Project: Introduction to Student Projects	Project Work: Preparation of Project Outlines
28.10.2020	Project: Feedback about Project Outlines	Exercise: Introduction to MapForce
04.11.2020	Coaching: Schema Mapping	Lecture: Identity Resolution
11.11.2020	Q&A: Identity Resolution	Exercise: Identity Resolution
18.11.2020	Coaching: Identity Resolution	Lecture: Data Quality and Data Fusion
25.11.2020	Q&A: Data Quality and Data Fusion	Exercise: Data Quality and Data Fusion
02.12.2020	Coaching: Data Quality and Fusion	Project Work: Data Quality and Fusion
09.12.2020	Presentation of project results	Presentation of project results
14.12.2020	Final Exam	

Details about Phase I: Data Collection and Data Translation

- Duration: now – November 11th
- Today
  1. Form teams of four people
  2. Decide on a domain/use case
  3. Start data collection and profiling

- Until Sunday, October 25th, 23:59
  - Send a 4 page abstract on your project (details next slide)
- Wednesday, October 28th,15:30-17:00
  - You get feedback on your abstract (if necessary)
- Thursday, October 29th
  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 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 25th, 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
movie	director	string/ list	dataset1, dataset3
movie	year	date	dataset2, dataset3, dataset4
…	…		…

Possible Use Cases for Student Projects

- Movies
  - budget, actors, directors, oscar nominiations...
- 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://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: 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)

- The integrated data set should contain *all* attributes

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

- The integrated data set should contain *all* attributes

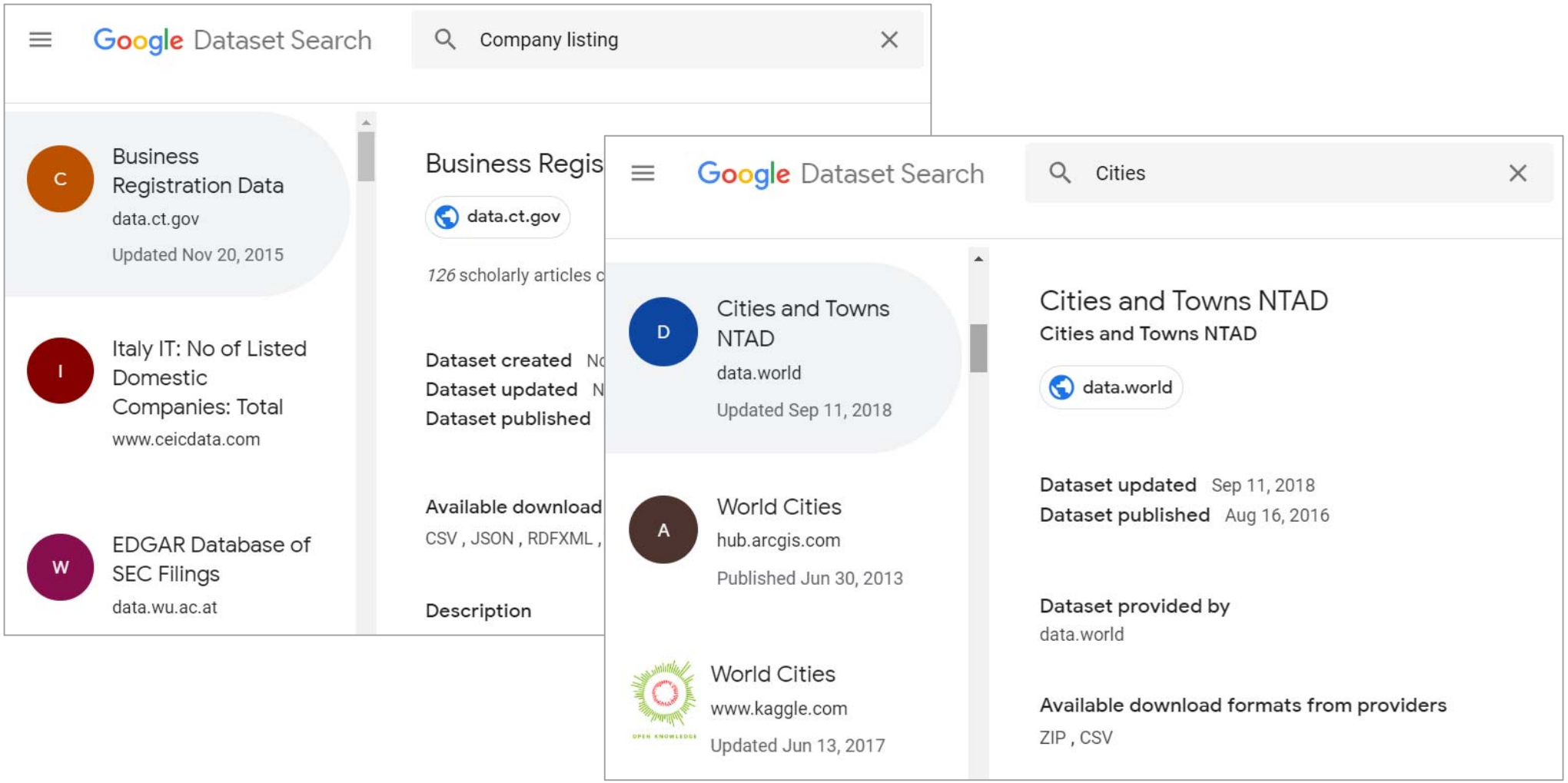
Example Use Case 3: Companies

- Forbes data set with top 2000 companies worlwide
  - 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")
 }
```

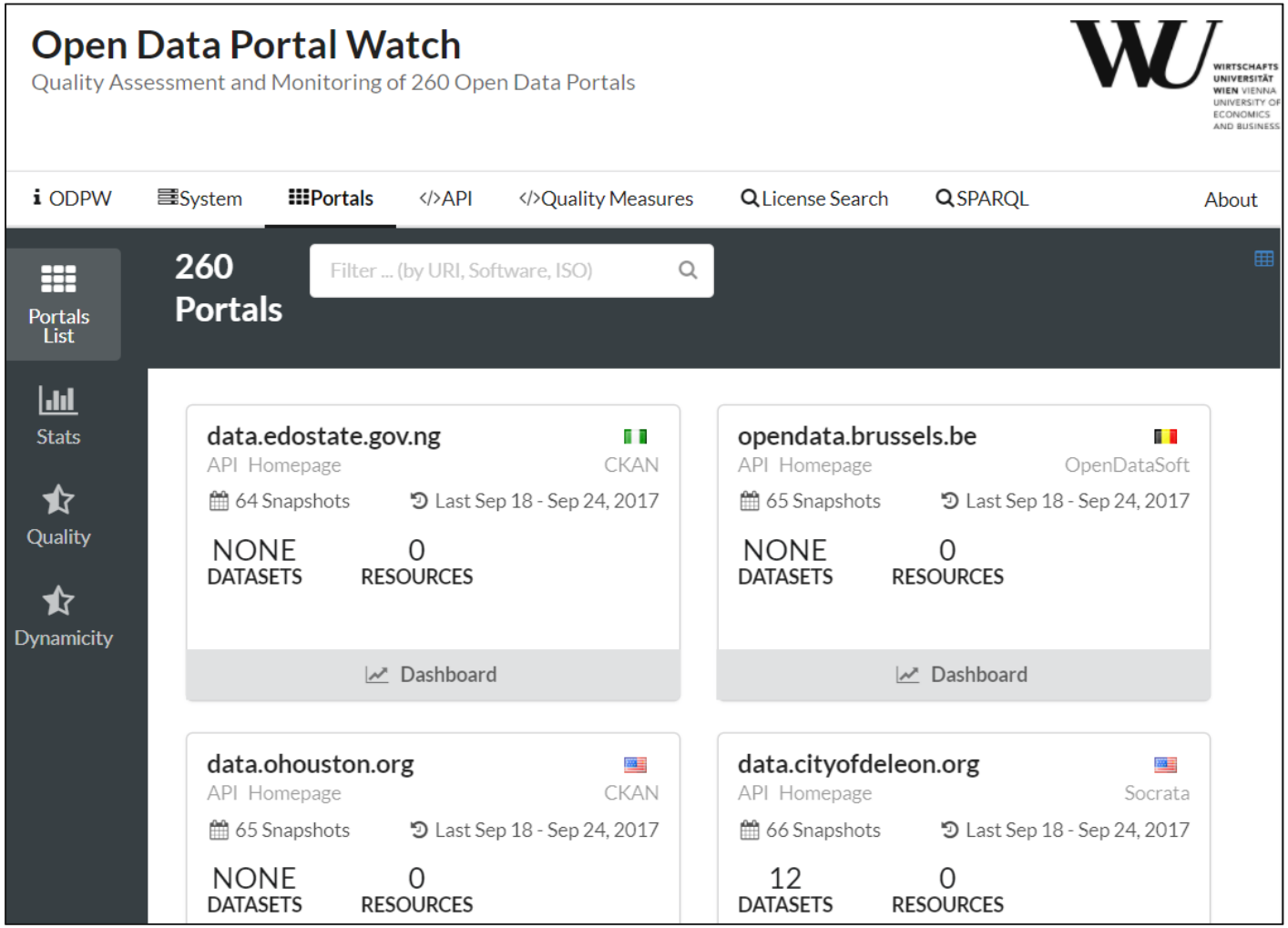
Where do I find Data for my Project?

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



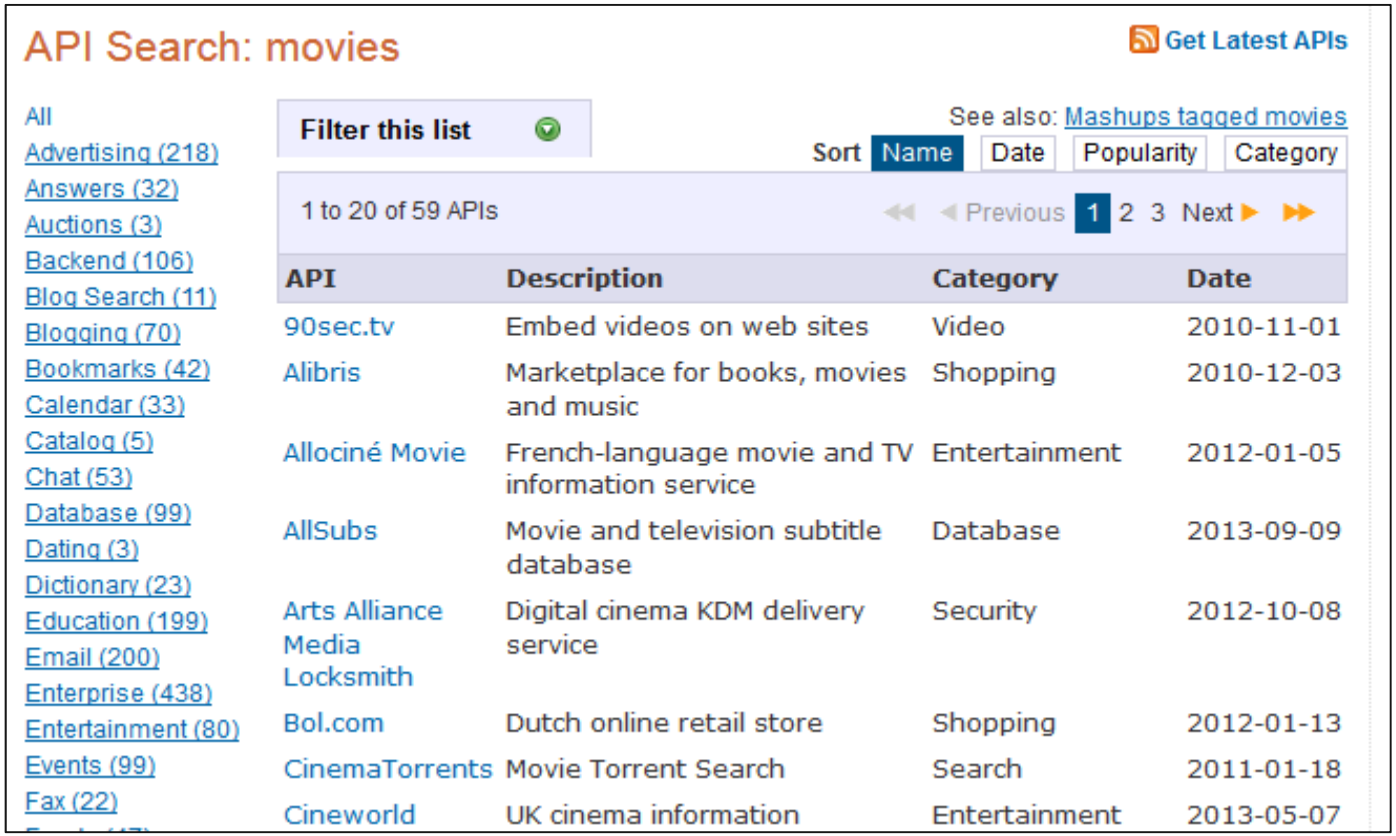
Where do I find Data for my Project?

- 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)



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

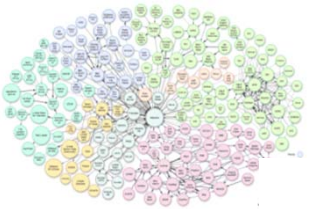
1 to 20 of 59 APIs

Previous 1 2 3 Next

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
- 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/

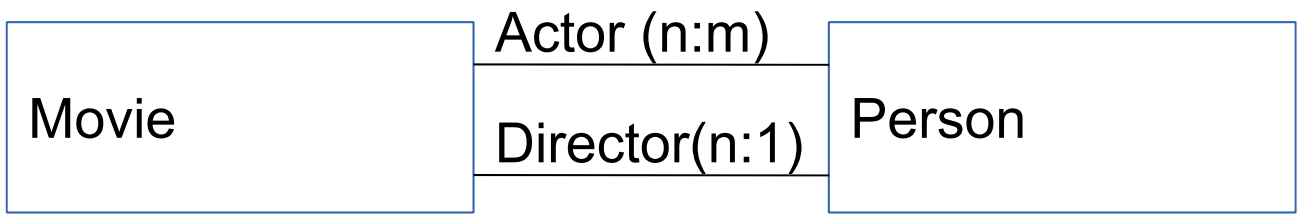
Class-Specific Subsets of the Schema.org Data

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...
   - Actor/Director: first name, last name, birth date, nationality, ...



Creating an Integrated Schema

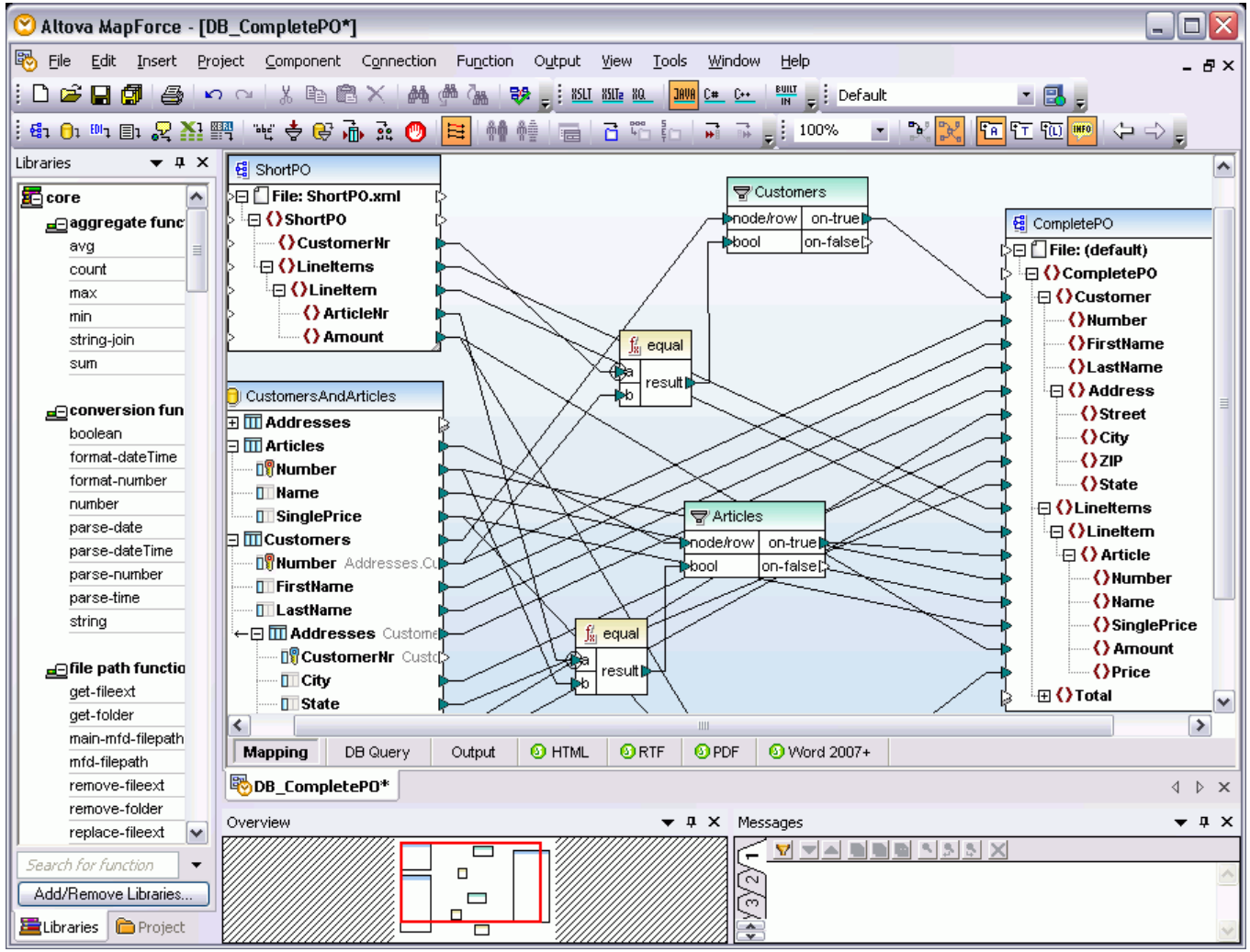
Hint: Create an example XML file

- 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 Anna and Alexander
2. Start translating your data into the unifying schema using MapForce



...and now

1. Team formation
   a. Students with team
      → Let us know and go ahead
   b. Students without team
      → Let us know and we will assign you
2. Agree on use case
3. Start collecting data