

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**

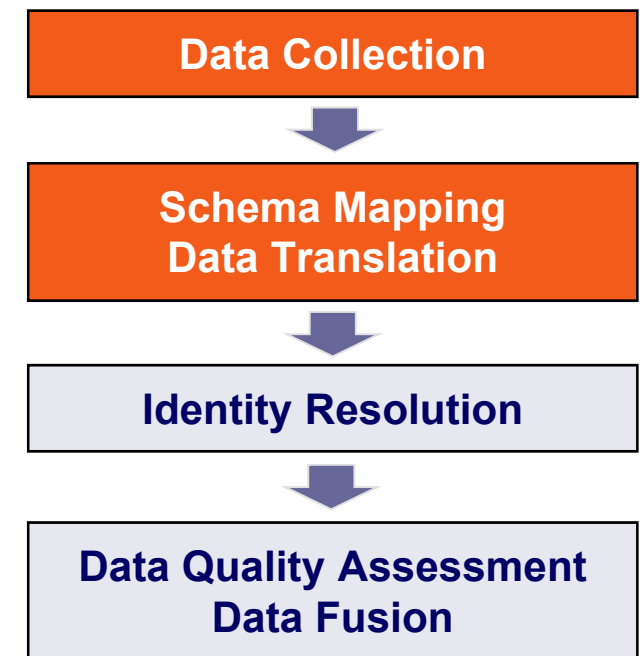
Duration: now – October 17th 2018

Tasks:

1. Find a partner (groups of five)
2. Decide on a use case
3. Collect data from the Web
4. Profile your data
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

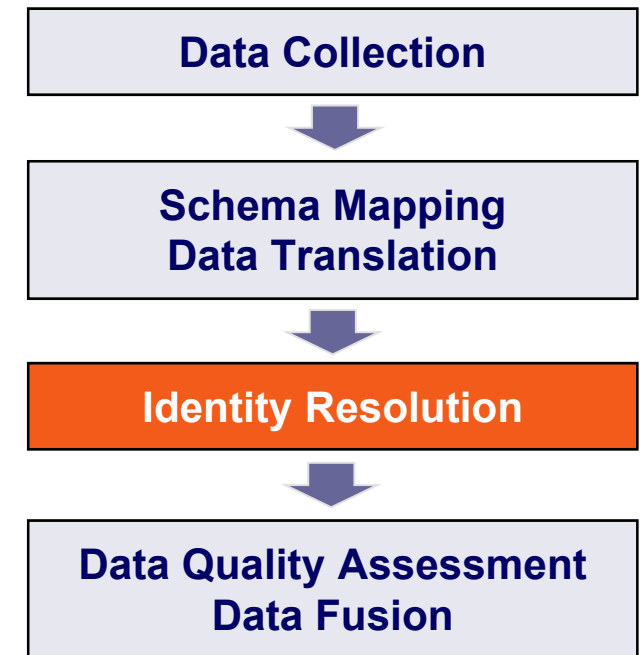
- **Phase II: Identity Resolution**

Duration: October 18th – November 7th

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**

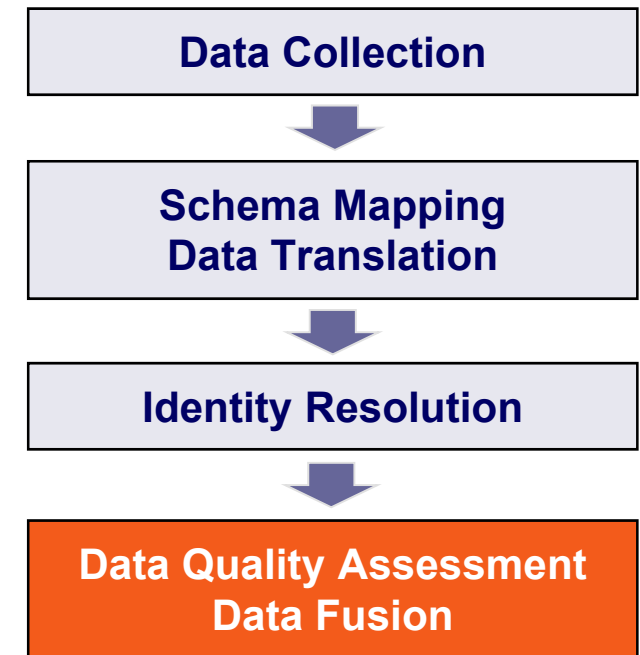
Duration: November 8th – December 1st

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 5th and December 6th
- Overview of your use case
- Explain your data
- Explain the strategies that you used
- Discuss the quality of your solution



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.

Details about Phase I: Data Collection and Data Translation

- Duration: now – October 17th 2017
- **Today**
 1. Form teams of four people
 2. Decide on a domain/use case
 3. Start data collection and profiling
- **Tomorrow**
 1. Introduction to **MapForce**
 2. Start using MapForce to translate data to target schema
- **Until Sunday, October 7th, 23:59**
 - Send a 4 page abstract on your project (details next slide)
- **Wednesday, October 10th**
 - You get feedback on your abstract



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 **2 classes** in target schema
 - e.g., movie and actor
4. at least **1000 entities** should be contained in at least **two datasets**
 - please estimate based on small sample
5. at least **10 attributes** in joint dataset
 - entities should be identifiable by attribute combination, e.g. name+birthdate
6. at least **5 attributes** should be contained in at least **two datasets**
 - some attributes should be contained in three datasets (for fusion by voting)
7. at least **2 different input file formats**
 - CSV, JSON, XML...

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 fulfil 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, Oliver Lehmberg and Christian Bizer
- **Deadline: Sunday, October 7th, 23:59**



Tables for 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, year, ...
DBpedia	dbpedia.org/sparql	xml	actor	23,500	8	name, birthDate (MV), activeYears,...
Freebase	Download URL	csv	actor	11,000	14	given_name, surname, spouse (MV)
.....						

(*) 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	Datasets in which attribute is found
movie	name	dataset1, dataset2, dataset3, dataset4
movie	director	dataset1, dataset3
movie	year	dataset2, dataset3, dataset4
.....		

Schedule

Week	Wednesday	Thursday
5.9.2018	Lecture: Introduction to Web Data Integration	Lecture: Structured Data on the Web
12.9.2018	Lecture: Data Exchange Formats	Lecture: Data Exchange Formats
19.9.2018	Lecture: Schema Mapping	Lecture: Schema Mapping
26.9.2018	Project: Introduction to Student Projects	Tool Intro: MapForce
3.10.2018	- Holiday -	Project Work: Data Translation
10.10.2018	Project: Feedback about Project Outlines	Lecture: Identity Resolution
17.10.2018	Lecture: Identity Resolution	Tool Intro: Winte.r Identity Resolution
24.10.2018	Coaching: Identity Resolution	Project Work: Identity Resolution
31.10.2018	Coaching: Identity Resolution	- Holiday -
7.11.2018	Lecture: Data Fusion	Lecture: Data Fusion
14.11.2018	Tool Intro: Winte.r Data Fusion	Project Work: Data Fusion
21.11.2018	Coaching: Data Fusion	Project Work: Data Fusion
28.11.2018	Coaching: Data Fusion	Project Work: Data Fusion
5.12.2018	Presentation of project results	Presentation of project results
17.12.2018	Final Exam	

Coaching Sessions

- Anna and Oliver will give you tips and answer questions concerning your project.
- Registration via email to Anna and Oliver is mandatory!
 - until Monday night!
 - including the questions that you like to discuss
- Oliver will assign you a time slot on Wednesday and inform you about the slot via email.

Possible Use Cases for Student Projects

- Movies and Actors
 - Actors, directors, budget, oscar nominations...
- Musicians and Bands
 - First name, last name, birth date, birth place, bands, albums ...
- Companies and Key Persons
 - company registries, panama papers data
- Statistical Data
 - Countries, regions, cities, population, area, leader, GDP, ...
- Books and Authors
 - 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. Actors and directors with birth dates



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

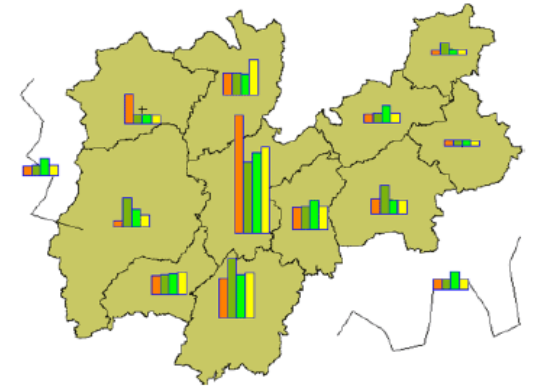


- Movie data from **DBpedia**
- Issue a SPARQL query against <http://dbpedia.org/sparql>
- Result can be stored as CSV, JSON, XML, ...

```
SELECT ?title ?budget ?gross ?director
WHERE {
    ?x a dbpedia-owl:Film .
    ?x dbpedia-owl:budget ?budget .
    ?x dbpedia-owl:gross ?gross .
    ?x dbpedia-owl:director ?d .
    ?d foaf:name ?director .
    ?x rdfs:label ?title .
    FILTER(LANG(?title)="en")
}
```

Example Use Case 2: Statistical Data

- Statistics about countries / regions / cities from CIA World Fact Book e.g., health data, transportation data, ...
 - <https://www.cia.gov/library/publications/the-world-factbook/rankorder/rankorderguide.html>
- European countries and regions:
 - <http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/>
- Country and city data from Geonames:
 - <http://download.geonames.org/export/dump/countryInfo.txt>
 - <http://download.geonames.org/export/dump/cities15000.zip>
- ...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



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 'Company listing'. The right screenshot shows search results for '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?



- **Google Table Search**

- <http://research.google.com/tables>
- Press
“Export to Fusion Tables”
- Press “See table”
- Use “File”->“Download”
to download a CSV

- **HTML Tables in General**

- e.g., Firefox plugin
“Export CSV”
- Exports a table to CSV
on right click

The screenshot shows the Google Tables search interface. At the top, the Google logo is on the left, and a search bar contains the word 'Cities' with a magnifying glass icon on the right. Below the search bar, the text 'Tables experimental' is on the left, and 'Ergebnisse 1 - 10 von ungefähr 171.709 für Cities (0.77 Sekunden)' is on the right. The main content area is divided into two columns. The left column has links for 'Web', 'Web Tables', 'Fusion Tables', and 'Send Feedback'. The right column displays search results. The first result is 'NYCdata: Top 100 Cities in the U.S. - Ranked by Population' with a URL. Below the link are buttons for 'Back', 'Top 100 Metropolitan ...', 'City', and 'New York', and a link 'Show less (106 rows / 5 columns total) - Export data'. Below this is a table with two buttons: 'Export to Google Sheets' and 'Export to FusionTables'. The table has five columns: 'Back', 'Top 100 Metropolitan', 'World Cities', 'World Cities', and 'Top 100 Metropolitan'. The first row of data shows 'Rank', 'City', 'State', 'Population', and 'Web Site'. The second row shows 'Notes: *Population' for each column. The third row shows 'View Additional' for each column. The fourth row shows 'Source: U.S. Census' for each column. The second result is 'List of cities in the United Kingdom - Simple English Wikipedia ...' with a URL. Below the link are buttons for 'City', 'English Cities', 'Bath sup 1 sup ...', 'Birmingham sup ...', and 'Bradford sup ...', and a link 'Show more (74 rows / 5 columns total) - Export data'.

Back	Top 100 Metropolitan	World Cities	World Cities	Top 100 Metropolitan
Rank	City	State	Population	Web Site
Notes: *Population	Notes: *Population	Notes: *Population	Notes: *Population	Notes: *Population
View Additional	View Additional	View Additional	View Additional	View Additional
Source: U.S. Census	Source: U.S. Census	Source: U.S. Census	Source: U.S. Census	Source: U.S. Census

Where do I find Data for my Project?

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

The screenshot displays the 'Open Data Portal Watch' website, which is dedicated to the 'Quality Assessment and Monitoring of 260 Open Data Portals'. The site features a dark navigation bar with links to ODPW, System, Portals, API, Quality Measures, License Search, SPARQL, and About. A sidebar on the left offers options for Portals List, Stats, Quality, and Dynamicity. The main content area, titled '260 Portals', includes a search filter and a grid of portal cards. Each card provides details for a specific data portal, including its name, country flag, data source (API or Homepage), snapshot count, last update period, dataset count, and resource count. A 'Dashboard' link is available for each portal.

Portal Name	Country	Source	Snapshots	Last Update	Datasets	Resources
data.edostate.gov.ng	🇳🇮	CKAN	64	Last Sep 18 - Sep 24, 2017	NONE	0
opendata.brussels.be	🇧🇪	OpenDataSoft	65	Last Sep 18 - Sep 24, 2017	NONE	0
data.ohouston.org	🇺🇸	CKAN	65	Last Sep 18 - Sep 24, 2017	NONE	0
data.cityofdeleon.org	🇺🇸	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** 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 dbpedia-owl:Film .
    OPTIONAL {?x dbpedia-owl: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

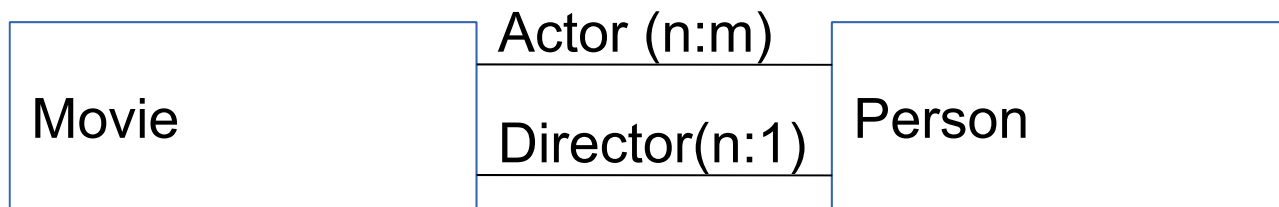
Class Name	Total Number of	Top Classes (Entity Count)	Total File Size	Quad File
http://schema.org/AdministrativeArea	Quads: 1,724,857 URLs: 85,625 Hosts: 63	http://schema.org/AdministrativeArea (100,671) http://schema.org/GeoCoordinates (84,152) http://schema.org/Country (83,851) http://schema.org/Continent (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 9**)
 - Create the tables for the project abstract (see **Slide 11**)
3. Apply schema integration method from lecture (Spaccapietra, et al.)

E.g.

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



Creating an Integrated Schema

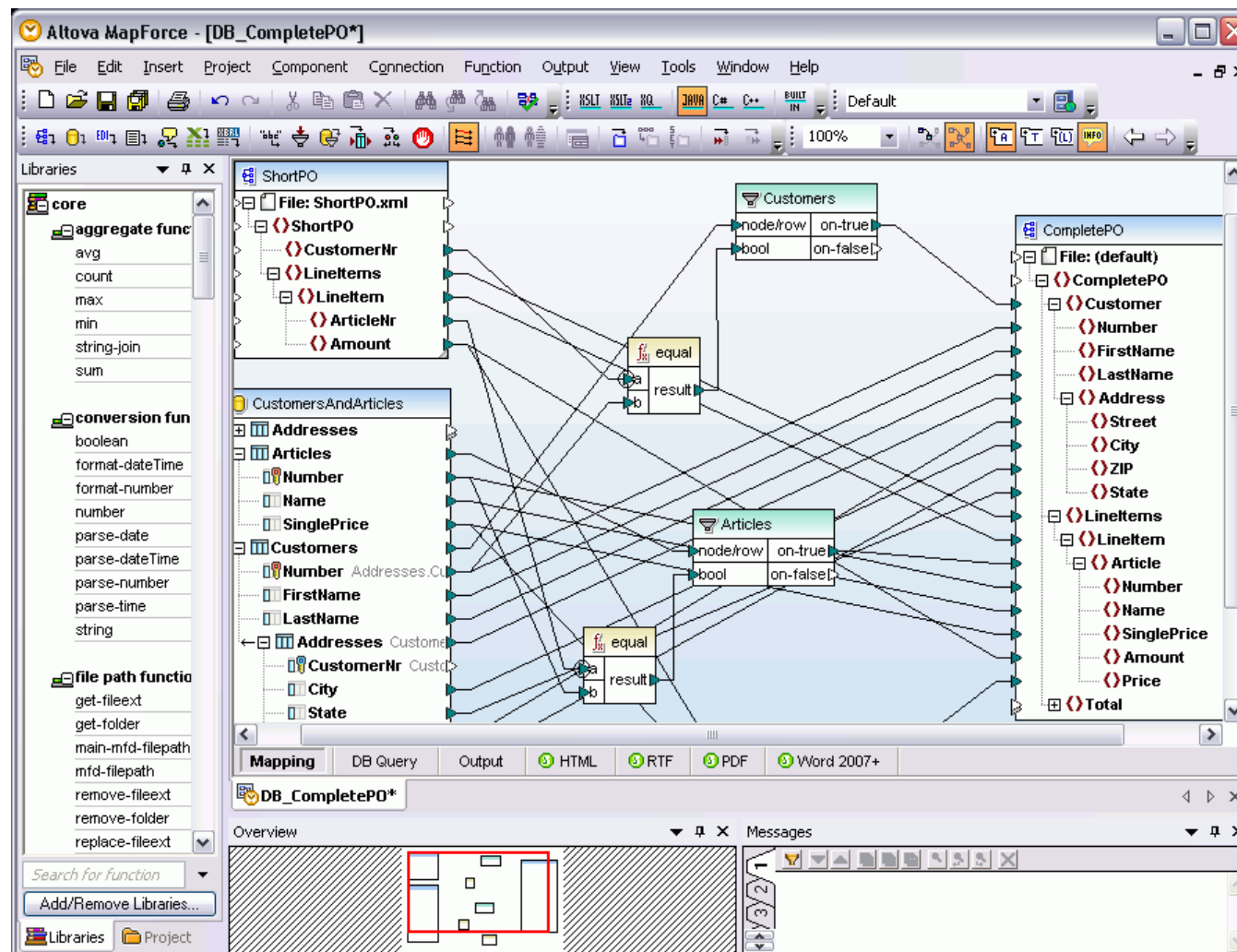
Hint: Create an example XML file

- 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

1. Introduction to MapForce
2. Translate your data into the unifying schema using MapForce



...and now

1. Find a partner
2. Agree on use case
3. Start collecting data

