UNIVERSITY OF MANNHEIM

Exercise: Data Fusion

Web Data Integration

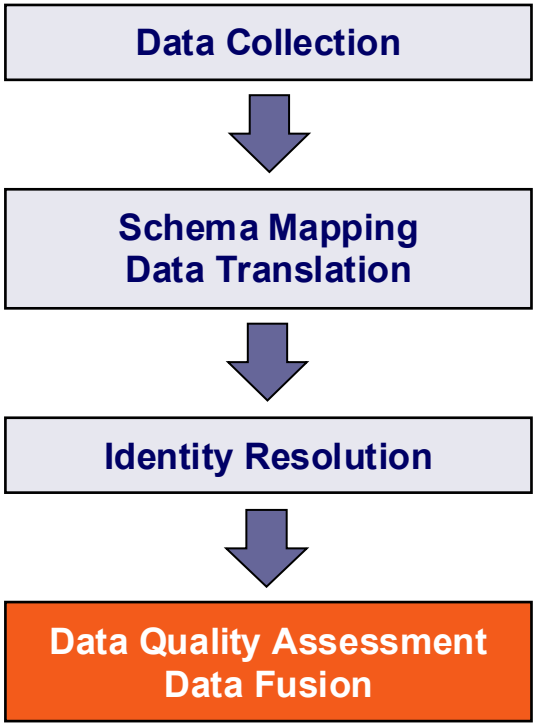


Exercise Overview

Learning goal

Learn how to use the PyDI Framework to:

1. Fuse data and resolve data conflicts
2. Experiment with different conflict resolution functions
3. Measure the density and accuracy of the resulting fused dataset



Result

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. A well-founded idea about the quality and profile of the fused data

The PyDI Framework

- The **Python Data Integration Framework** (PyDI) provides methods for end-to-end data integration
- Implements methods for
  - Data Pre-Processing
  - Schema Matching
  - Identity Resolution
  - **Data Fusion**
  - Evaluation

- Open Source under Apache 2.0 License
- https://github.com/wbsg-uni-mannheim/PyDI

Provided Datasets and Correspondences

academy_awards_2_actors_correspondences.csv

```
49  academy_awards_1880,actors_49,1.0
50  academy_awards_1829,actors_50,1.0
51  academy_awards_1776,actors_51,1.0
52  academy_awards_1731,actors_52,1.0
53  academy_awards_1680,actors_53,1.0
```

(i) academy awards – actors correspondences

actors_2_golden_globes_correspondences.csv

```
28  actors_49,golden_globes_1733,1.0
29  actors_50,golden_globes_1684,1.0
30  actors_51,golden_globes_1656,1.0
31  actors_52,golden_globes_1608,1.0
32  actors_53,golden_globes_1566,1.0
```

(ii) actors – golden globes correspondences

```
<movie>
    <id>academy_awards_1880</id>
    <title>One Flew over the Cuckoo's Nest</title>
    <director>
        <name>Milos Forman</name>
    </director>
    <actors>
        <actor>
            <name>Jack Nicholson</name>
        </actor>
        <actor>
            <name>Brad Dourif</name>
        </actor>
        <actor>
            <name>Louise Fletcher</name>
        </actor>
    </actors>
    <date>1975-01-01</date>
    <oscar>yes</oscar>
</movie>
```

(i) academy awards dataset

metadata:

dataset_score: 0.4

dataset_date: 2015

```
<movie>
    <id>golden_globes_1733</id>
    <title>One Flew Over The Cuckoo''s Nest</title>
    <director>
        <name>Milo Forman</name>
    </director>
    <actors>
        <actor>
            <name>Louise Fletcher</name>
        </actor>
        <actor>
            <name>Jack Nicholson</name>
        </actor>
    </actors>
    <date>1976-01-01</date>
    <globe>yes</globe>
</movie>
```

(ii) golden globes dataset

metadata:

dataset_score: 0.5

dataset_date: 2018

```
<movie>
    <id>actors_49</id>
    <title>One Flew Over The Cockoo□s Nes</title>
    <actors>
        <actor>
            <name>Louise Fletcher</name>
            <birthday>1934-01-01</birthday>
            <birthplace>Alabama</birthplace>
        </actor>
    </actors>
    <date>1976-01-01</date>
</movie>
```

(iii) actors dataset

metadata:

dataset_score: 0.6

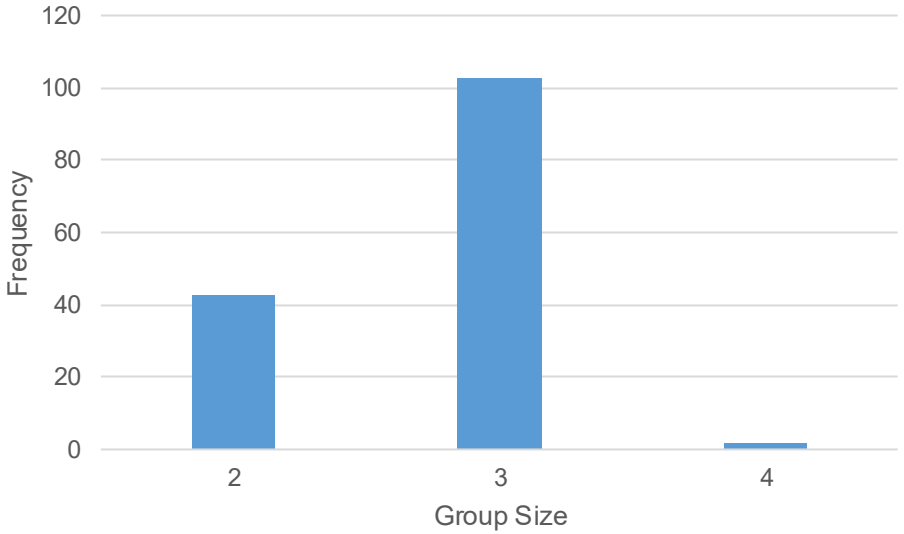
dataset_date: 2018

Profile Correspondences: Group Size Distribution

- Check group size distribution
  - All correspondences from the identity resolution are combined
  - All records that are believed to describe the same real-world entity end up in a group (transitive!)

```
Group Size Distribution of 148 groups:
Group Size | Frequency
———————————————————————
2          |          43
3          |          103
4          |          2
```

Due to matching errors
or duplicates in sources



Evaluating Fusion: We need an evaluation set!

Record generated by data fusion

```
<movie>
  <id>academy_awards_1880+actors_126+actors_49+golden_globes_1733</id>
  <title provenance="golden_globes_1733">One Flew Over The Cuckoo''s Nest</title>
  <director provenance="academy_awards_1880">Milos Forman</director>
  <date provenance="academy_awards_1880">1975-01-01T00:00</date>
  <actors provenance="actors_126+academy_awards_1880+golden_globes_1733+actors_49">
    <actor>
      <name>Brad Dourif</name>
    </actor>
    <actor>
      <name>Louise Fletcher</name>
    </actor>
    <actor>
      <name>Jack Nicholson</name>
    </actor>
  </actors>
</movie>
```

title with the longest string

director from most trusted source

actors list from union

Evaluation set record

```
<movie>
    <id>academy_awards_1880</id>
    <title>One Flew over the Cuckoo's Nest</title>
    <director>
        <name>Milos Forman</name>
    </director>
    <actors>
        <actor>
            <name>Jack Nicholson</name>
        </actor>
        <actor>
            <name>Brad Dourif</name>
        </actor>
        <actor>
            <name>Louise Fletcher</name>
        </actor>
    </actors>
    <date>1975-01-01</date>
    <oscar>yes</oscar>
</movie>
```

Compare the fused values with the ones in the evaluation set:

- Did your conflict resolution functions work?
- Should the value comparison be strict or more relaxed?
  - „Cuckoo‘‘s Nest“ *vs* „Cuckoo‘s Nest“

PyDI Tutorial also covers Data Fusion

https://github.com/wbsg-uni-mannheim/PyDI/blob/main/docs/tutorial/entity_matching_and_fusion/data_integration_tutorial.ipynb

PyDI Data Integration Tutorial

This tutorial demonstrates comprehensive data integration using PyDI. We'll work with movie datasets to showcase the data integration pipeline from entity matching to Data Fusion.

What You'll Learn

1. **Data Loading & Profiling**: Load and analyze movie datasets with provenance tracking
2. **Entity Matching**:
   - Blocking strategies (Standard, Sorted Neighbourhood, Token-based, Embedding-based)
   - Multi-attribute similarity matching with custom comparators
   - Machine learning-based entity matching
3. **Data Fusion**:
   - Conflict resolution with custom fusion rules
   - Quality assessment against test set
   - Provenance-based conflict resolution

Datasets

We'll use three movie datasets:

- **Academy Awards**: Movies with Oscar information (4,592 records)
- **Actors**: Movies with actor details (149 records)
- **Golden Globes**: Movies with Golden Globe awards (2,286 records)

These datasets contain overlapping movie information but with different attributes, data quality issues, and conflicting values - perfect for demonstrating real-world data integration challenges.

In [1]:

```
from pathlib import Path

# Setup paths
def get_repo_root():
    """Get repository root directory."""
    current = Path.cwd()
    while current != current.parent:
        if (current / 'pyproject.toml').exists():
            return current
        current = current.parent
    return Path.cwd()
```

Project Related Information

Select Data for Fusion Experiments

- Your input is the output of Exercises 1 and 2
  - schema is aligned, unique IDs are in place, identity resolution is done
- What to look out for during fusion experiments?
  1. Make sure again that attribute intersection is big enough
     - you should be able to fuse data for >= 5 attributes
  2. Make sure that the quality of identity resolution is good enough
  3. it makes sense to apply different fusion functions to the attributes
     - compare different functions
     - experiment with using provenance data vs. no provenance data

Prepare Your Evaluation Set

- Your evaluation set should contain
  - >= 15 entities
  - >= 4 attributes per entity
- Manually look up the correct values (in an external data source)!
- The evaluation set uses your target schema
  - Use IDs from one of your data sets, so the evaluator can find the correct records!
- Don't create ambiguities!
  - If records a, b, c are the same according to the identity resolution
  - Only one corresponding record can appear in the evaluation set
  - Choose a, b or c as ID in the evaluation set

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 on Sunday, 29th November 2026, 23:59
  - Submit via ILIAS
- You must use latex template of Springer CS Proceedings
  - https://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines
- Also **submit**
  1. your code and
  2. (a subset) of your data
- Please cite sources properly if you use any
  - Use citation style of the latex template!

Final Report: Content

Your final report should contain

1. Results of Phase 1: Data Translation
2. Results of Phase 2: Identity Resolution
3. Results of Phase 3: Data Fusion
4. Summary of the overall result

Data Translation in the Final Report

- Your report should contain
  1. Profiling results describing your input data sets
     - e.g. updated versions of the tables that you created for your project proposal

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)
.....						

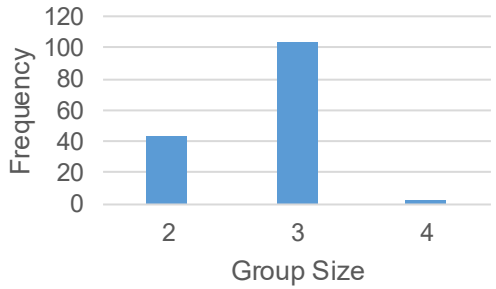
Class name	Attribute name	Datasets in which attribute is found
movie	name	dataset1, dataset2, dataset3, dataset4
movie	director	dataset1, dataset3
movie	year	dataset2, dataset3, dataset4
.....		

  2. Your consolidated schema and how you created it
  3. Which transformations you used and why
     - if there was any information you could not transform

Identity Resolution in the Final Report

- Your report should contain
  1. Content and size of your evaluation sets and procedure used to create them
  2. Which matching rules/matchers did you try?
     - Discuss what happened with P/R and F1?
     - Please include a table comparing the results of the different matchers
  3. Which blocking methods did you try?
     - Report and discuss the change in runtime, number of matches, and reduction ratio. How do P/R/F1 change?
     - Please include a table comparing the results of the different blocking methods that you tested
  4. What's the group size distribution of your result?
  5. An analysis of the errors that remain when applying your best matching rule.

#	Matcher	Blocker	P	R	F1	# Corr	Time
1	Rule1:Title&Year	No Blocking	0.71	0.95	0.82	10.230	90 min
2	Rule1:Title&Year	StandardYear	0.71	0.73	0.72	9.609	18 sec
3	Rule1:Title&Year	SNBYear	0.71	0.89	0.79	10.215	50 sec
4	Rule2:Title&Actors	SNBYear	0.81	0.89	0.83	9.919	19 sec



Data Fusion in the Final Report

- Your report should contain
  1. What kind of provenance data you added?
  2. What was the density of your input and the merged datasets?
  3. How consistent were your datasets?
  4. Size and content of your evaluation set and how you created it
  5. Which conflict resolution functions you tried for each attribute
     - Whether you defined your own conflict resolution functions
  6. Which accuracy did the different conflict resolution functions deliver? What was the best function for each attribute?
     - Please include a table comparing the accuracies reached by the different resolution functions.
  7. An analysis of the errors that remain after applying your best resolution function.

Summary of Overall Results in Final Report

Please conclude your report with a summary answering the following three questions:

1. How many additional entities did you add compared to the largest of your input datasets?
2. How much did you increase the density of your data compared to the largest of your input datasets?
3. What is the overall accuracy of your final dataset according to your evaluation set?

Final Report: Important

- Balance your content between the 3 exercises
  - not 10 pages on identity resolution and 2 pages on the rest
- If you have done something cool – write about it!
  - it is highly unlikely we dig it out of your code ourselves
- We are strict about the 12 pages limit. Thus,
  - include lots of tables to show us what you have tried
  - briefly discuss the results of each of your experiments
  - do not repeat theoretical stuff from the slides (e.g. definition of X)
  - we will reduce your mark by 0.33 for each extra page
    no matter how interesting it is!

Final Presentation and Exam

- Presentation Dates
  - 02.12.2026 and 03.12.2026
- Date of WDI **offline** Exam
  - 17.12.2026
- Format
  - 5-6 open questions that show that you have understood the theory part of the lecture
    - all lecture slide sets including structured data on the Web and data exchange formats + query languages XPath and SPARQL
  - Duration: 60 minutes

...and now

- Take your IR outputs and
  - Experiment with creating the fused dataset, and density and consistency evaluation metrics
  - Define your conflict resolution functions
  - Define your evaluation set
  - Experiment with data fusion and accuracy evaluation metrics
- **Write your final report**
- **Repeat the theory parts in order to be ready for the final exam**

