



# Information Integration Course Project

Thorsten Papenbrock  
WS 2015 / 2016

# Information Integration Course Project – Tasks

## Task 1: Extraction

Task 2: Integration Planning

Task 3: Integration Execution

Task 4: Cleansing

Task 5: Visualization

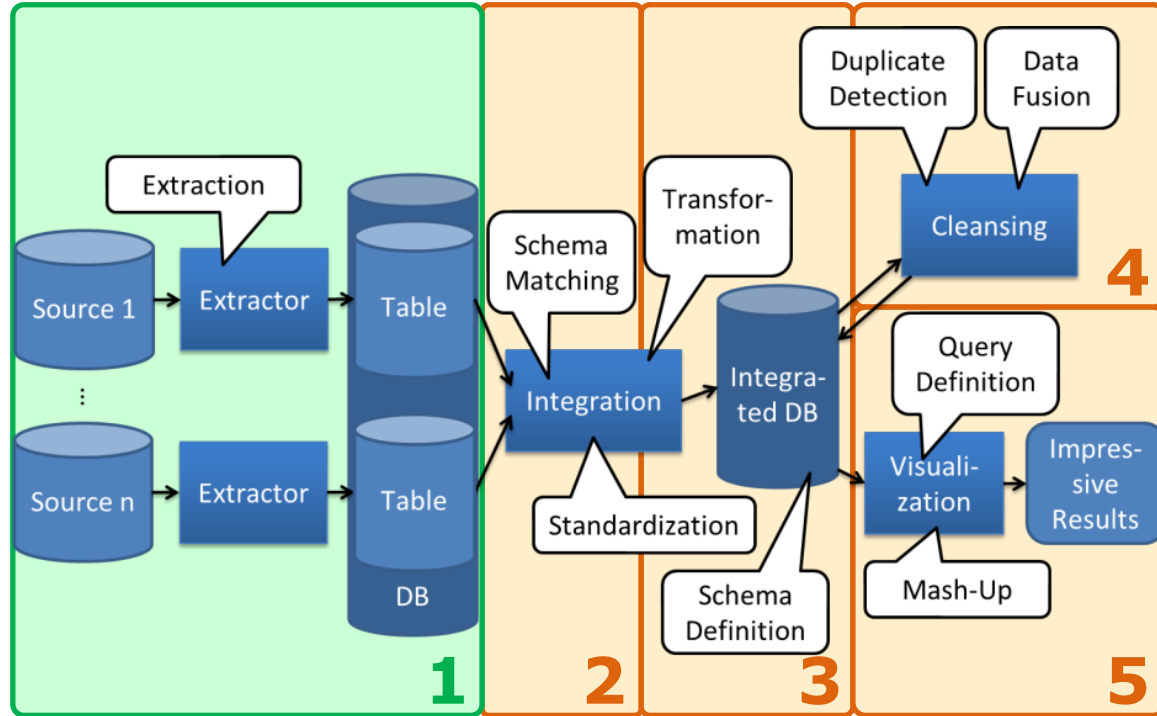
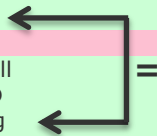


Chart 2

1	GND		<a href="http://datendienst.dnb.de">http://datendienst.dnb.de</a>
2	Baseball archive		<a href="http://seanlahman.com/baseball-archive/statis">http://seanlahman.com/baseball-archive/statis</a>
3	Abgeordnetenwatch		<a href="https://www.abgeordnetenwatch.de/api">https://www.abgeordnetenwatch.de/api</a>
4	DBpedia		<a href="http://dbpedia.org">dbpedia.org</a>
5	MusicBrainz		<a href="https://musicbrainz.org">https://musicbrainz.org</a>
6	IMDb		<a href="http://www.imdb.com/">http://www.imdb.com/</a>
	Freebase		<a href="http://wiki.freebase.com/wiki/Main_Page">http://wiki.freebase.com/wiki/Main_Page</a>
	WordNet		<a href="https://wordnet.princeton.edu/wordnet">https://wordnet.princeton.edu/wordnet</a>
	Discogs		<a href="http://www.discogs.com">http://www.discogs.com</a>
7	Bundesliga		<a href="http://dbup2date.uni-bayreuth.de/bundesliga.html">http://dbup2date.uni-bayreuth.de/bundesliga.html</a>
8	Wikidata		<a href="https://www.wikidata.org">https://www.wikidata.org</a>
9	Whitehouse		<a href="https://open.whitehouse.gov">https://open.whitehouse.gov</a>
10	SoccerWM2014		<a href="https://docs.google.com/spreadsheets/d/1i7aUrbCcle7">https://docs.google.com/spreadsheets/d/1i7aUrbCcle7</a>
	OpenFootball		<a href="https://github.com/openfootball">https://github.com/openfootball</a>
11	Judobase		<a href="https://www.judobase.org">https://www.judobase.org</a>
12	BasketballValue		<a href="http://basketballvalue.com/downloads.php">http://basketballvalue.com/downloads.php</a>
13	TheMovieDB		<a href="https://www.themoviedb.org/">https://www.themoviedb.org/</a>
14	databaseBasketball		<a href="http://www.databasebasketball.com">http://www.databasebasketball.com</a>
15	Ergast		<a href="http://ergast.com/">http://ergast.com/</a>
16	Open Library		?
	SoccerWiki		<a href="http://en.soccerwiki.org/">http://en.soccerwiki.org/</a>
17	Hockey		<a href="http://www.opensourcesports.com/hockey/">http://www.opensourcesports.com/hockey/</a>
18	Movies		<a href="http://kdd.ics.uci.edu/databases/movies/movies.html">http://kdd.ics.uci.edu/databases/movies/movies.html</a>
19	omdb		<a href="http://www.omdb.org/content/Help/DataDownload">http://www.omdb.org/content/Help/DataDownload</a>
	Football		<a href="http://www.databasefootball.com">http://www.databasefootball.com</a>
20	Basketball		<a href="http://www.opensourcesports.com/basketball">http://www.opensourcesports.com/basketball</a>
21	UCI_KDD		<a href="http://kdd.ics.uci.edu/databases/movies">kdd.ics.uci.edu/databases/movies</a>
22	omdb.org		omdb.org
23	Cinemalytics		cinemalytics.com
24	Bundestag		<a href="http://www.bundestag.de/abgeordnete">http://www.bundestag.de/abgeordnete</a>
25	Bundesrat		<a href="http://www.bundesrat.de/DE/bundesrat/mitglieder/...">http://www.bundesrat.de/DE/bundesrat/mitglieder/...</a>
26	FootballProject		<a href="http://http://footballproject.com">http://http://footballproject.com</a>
27	FootballData		<a href="http://api.football-data.org/">http://api.football-data.org/</a>
28	NBA BIOGRAPHICAL DATABASE		<a href="http://www.apbr.org/NBAData1.xls">http://www.apbr.org/NBAData1.xls</a>
	Bundesanzeiger		<a href="https://www.bundesanzeiger.de">https://www.bundesanzeiger.de</a>
	Angellist		<a href="https://angel.co/">https://angel.co/</a>
29	New York Times Linked Open Data		<a href="http://data.nytimes.com/">http://data.nytimes.com/</a>
30	Tennis Rankings		<a href="https://github.com/JeffSackmann/tennis_atp">https://github.com/JeffSackmann/tennis_atp</a>



# Information Integration Course Project – Tasks

Task 1: Extraction

**Task 2: Integration Planning**

Task 3: Integration Execution

Task 4: Cleansing

Task 5: Visualization

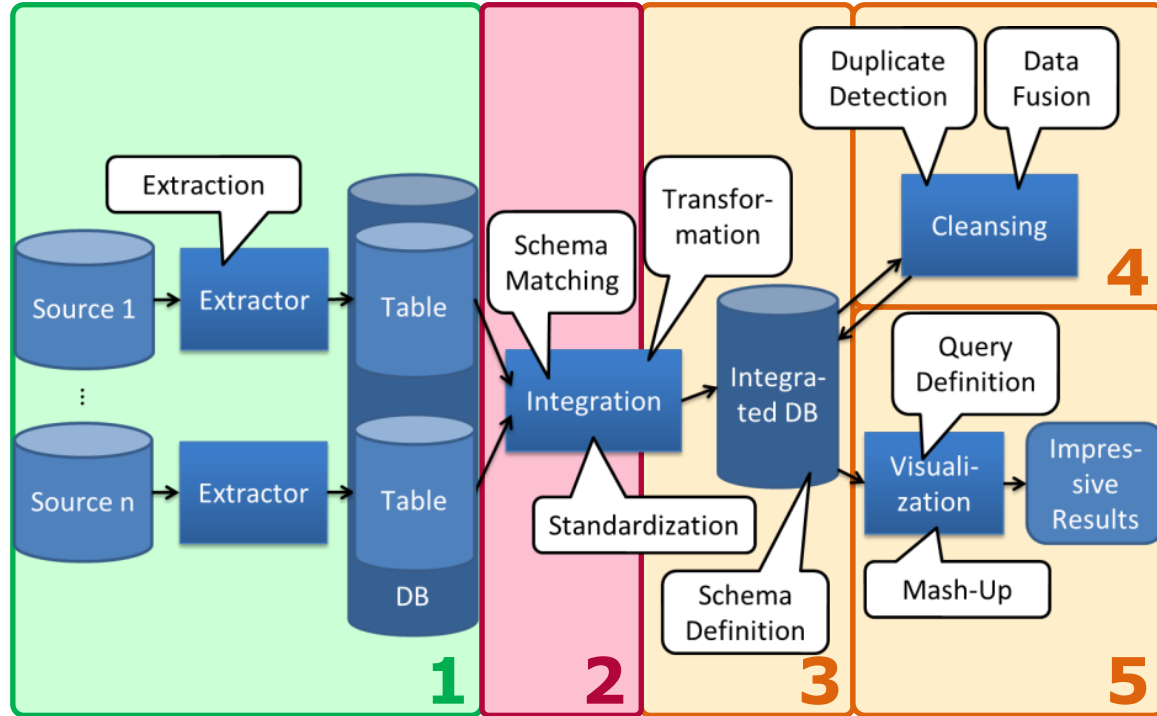


Chart 4

# Information Integration

## Course Project – Integration Planning

### 1. Export your data for your class mates:

- Create a PostgreSQL sql dump that creates your databases:  
→ include schemata and data!

```
CREATE DATABASE infointe WITH TEMPLATE = template0 ENCODING = 'UTF8' LC_COLLATE = 'de_DE.UTF-8' LC_CTYPE = 'de_DE.UTF-8';

ALTER DATABASE infointe OWNER TO postgres;

\connect infointe

CREATE SCHEMA public;

ALTER SCHEMA public OWNER TO postgres;

CREATE EXTENSION IF NOT EXISTS plpgsql WITH SCHEMA pg_catalog;

COMMENT ON EXTENSION plpgsql IS 'PL/pgSQL procedural language';

SET search_path = public, pg_catalog;
SET default_tablespace = '';
SET default_with_oids = false;
```

```
CREATE TABLE club (
    id integer NOT NULL,
    "Name" character varying(50),
    "Liga" character varying(50)
);

ALTER TABLE ONLY club ALTER COLUMN id SET DEFAULT nextval('club_id_seq'::regclass);

COPY club (id, "Name", "Liga") FROM stdin;
1 Botafogo Brazil
2 FC Barcelona Spain
3 Paris Saint-Germain France
4 Chelsea England
5 Manchester City England
...
```

- Name each dump according to the datasets it contains:  
<datasetA>\_<datasetB>.sql
- Copy your dump(s) to:  
\\fs23\bbs\Studentenaustauschordner\InfoInt2015\Base Schemata

**Information  
Integration  
Project**

Thorsten Papenbrock,  
WS 2015 / 2016

Chart 5

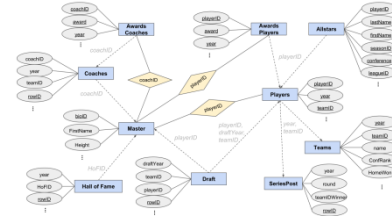
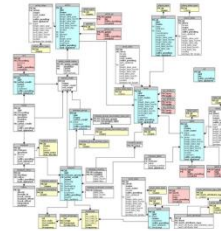
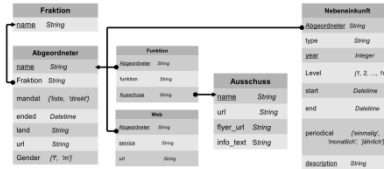
# Information Integration Course Project – Integration Planning

## 2. Review the datasets from the other teams:

- Get an overview of all datasets.



- Collect the ER-Diagrams.



- If explanations are necessary:
  - Use the mailing list or contact the authors directly.



**Information  
Integration  
Project**

Thorsten Papebrock,  
WS 2015 / 2016

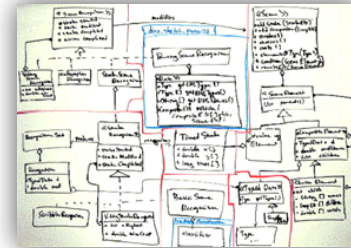
Chart 6

# Information Integration

## Course Project – Integration Planning

### 3. Develop your integrated schema:

- a) Define a topic for your integrated schema, for example:
  - “Athletes”, “Artists”, “Government”, “Media Industry”, “Celebrities”, ...
- b) Gather all schemata that suite your topic (must not be your own schemata):
  - Use at least 8 schemata and 2 relations from each schema.  
→ at least 16 base schemata.
- c) Merge equivalent relations and attributes.
  - Keep as much information as possible.
- d) Consolidate and normalize the integrated schema.
  - Create new relations that arise from different attributes if necessary.
  - Link relations from different base schemata if possible.
  - Try to avoid redundancy and NULL values via normalization.



### Information Integration Project

Thorsten Papenbrock,  
WS 2015 / 2016

Chart 7

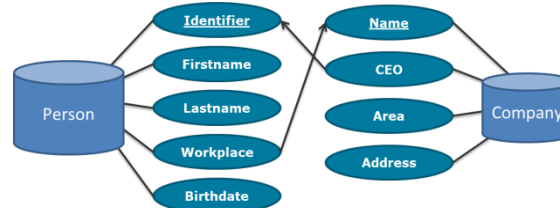
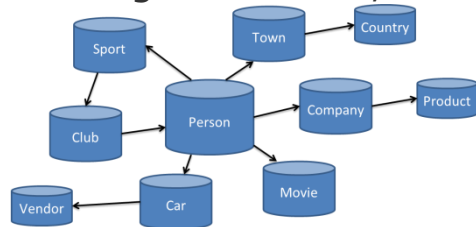
Note: You do not need to execute the integration for your datasets yet!

# Information Integration

## Course Project – Integration Planning

### 4. Document your integrated schema:

- 1) Manifest the integrated schema in a new PostgreSQL database instance:
  - Create all tables and their key and foreign-key constraints.
  - Export the empty(!) database instance as an SQL script and copy it to:  
[\\fs23\bbs\Studentenaustauschordner\InfoInt2015\Integrated\\_Schemata](#)
- 2) Describe the integrated schema in your presentation slides:
  - Introduce and motivate the topic of your integrated schema.
  - Explain your integrated schema using an ER- or class-diagram.  
→ For large schemata, use different abstraction layers.



### Information Integration Project

Thorsten Papenbrock,  
WS 2015 / 2016


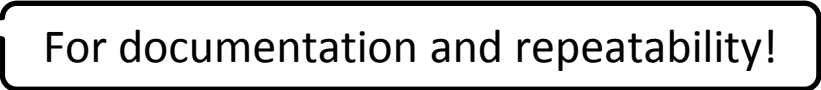
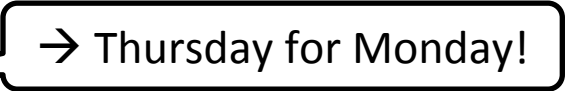
Chart 8

- Describe what data(sets) you have integrated.
- Discuss the problems that you encountered and the tools you used.



# Information Integration

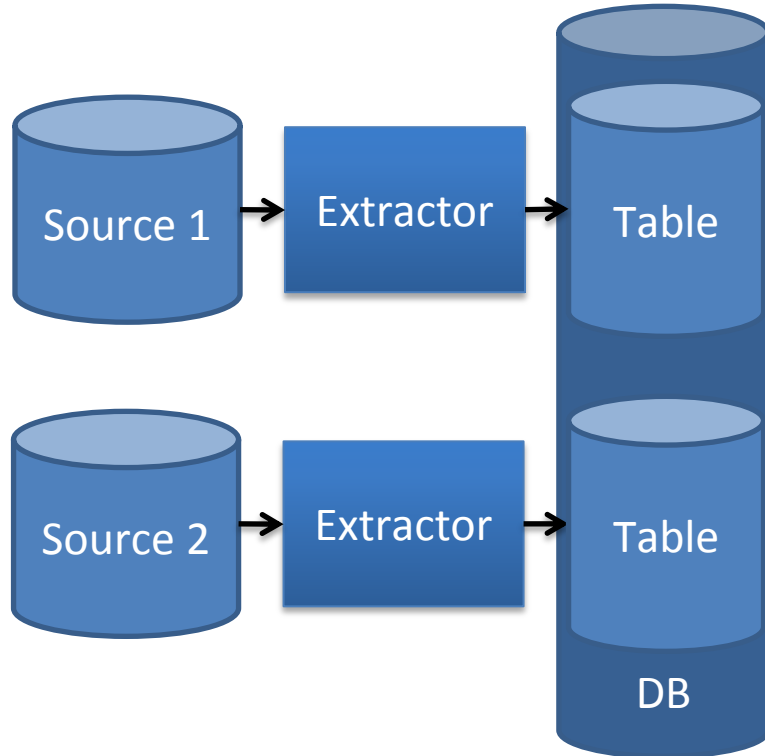
## Course Project – Deliverables

- **3-5 Slides** 
  - for **<5 min** presentations in class
  - in **pdf** format
  - with name  
**<last-name1>\_<last-name2>\_<last-name3>\_<last-name4>.<pdf>**
- **Database dumps** 
- *Submission:*
  - *Channel: Email at [thorsten.papenbrock\(at\)hpi.de](mailto:thorsten.papenbrock@hpi.de)*
  - *Subject: **[InfoInt2015] Exercise <NR> <last-name1>***
  - *Deadline: **Two work days before exercise lectures***   
*→ Monday for Wednesday lectures*

Note: Do not forget the **author names** on your slides!

→ Thorsten Papenbrock,  
WS 2015 / 2016

# Information Integration Presentations on Extraction



**Information  
Integration  
Project**

Thorsten Papenbrock,  
WS 2015 / 2016

Chart **10**

# Information Integration

## Course Project – Extraction Presentation

### Agenda:

1. Barkowsky Dumke Ihde Montenegro
2. Beck Duecker Mattfeld Schneider
3. Bleifuss Buelow Draeger Wong
4. Braatz Dinger Oldag Sauer
5. Eckert Gläser Hegner Zabel
6. Jasper Moritz Rzepka Werkmeister
7. Kliem Petrykowski Rehfeldt
8. Neuschäfer-Rube Pollak Reschke Rückert
9. Burmeister Zöllner
10. Frahnw Grundke Lindemann Sachse
11. Koall Stengel Mang Hempfing
12. Lange Martin Schirmer Walther
13. Perchyk Maschler Djürken Risch
14. Stamm Keßler Kimmig Kroschk

Advertise your data  
(and methods)!

### Information Integration Project

Thorsten Papenbrock,  
WS 2015 / 2016

Chart **11**



# Information Integration Course Project

Questions to:

Mailing List: [Infoint-2015@hpi.uni-potsdam.de](mailto:Infoint-2015@hpi.uni-potsdam.de)

Thorsten Papenbrock: Email or Office E-2-01.2