

Enrichment of contact data using information from public APIs

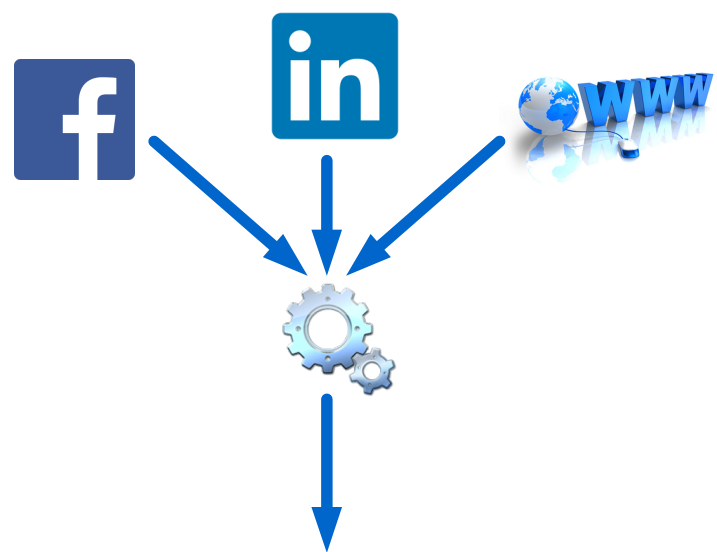
Markus Ehringer, 23th April 2018, Advanced Seminar Course for Graduands and Postgraduates

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

- Introduction
- Research Questions
- Sources analysis
- Prototype
- Data quality assessment
- Synchronization approach
- Demo

- Situation: Incomplete contact data

First Name	Last Name	Company	Age	E-Mail
Nick	Pisa	Google		nick.pisa@gmail.com
Robert	Tuner	BMW	45	
Michelle	Rosa	TUM	32	mr86@tum.de



First Name	Last Name	Company	Age	E-Mail
Nick	Pisa	Google	23	nick.pisa@gmail.com
Robert	Tuner	BMW	45	rob.tuner@gmx.at
Michelle	Rosa	TUM	32	mr86@tum.de

- 1) Which online sources could provide public available information for automatic processing?**
- 2) How can the quality of sources' information/data be assessed?**
- 3) How can the contact data, once enriched, be synchronized with the sources efficiently?**

Sources analysis

- List of 27 contacts → Google Search
- Sources:
 - Facebook
 - Xing
 - LinkedIn
 - Twitter
 - Crunchbase
 - AngleList
 - F6S

Result

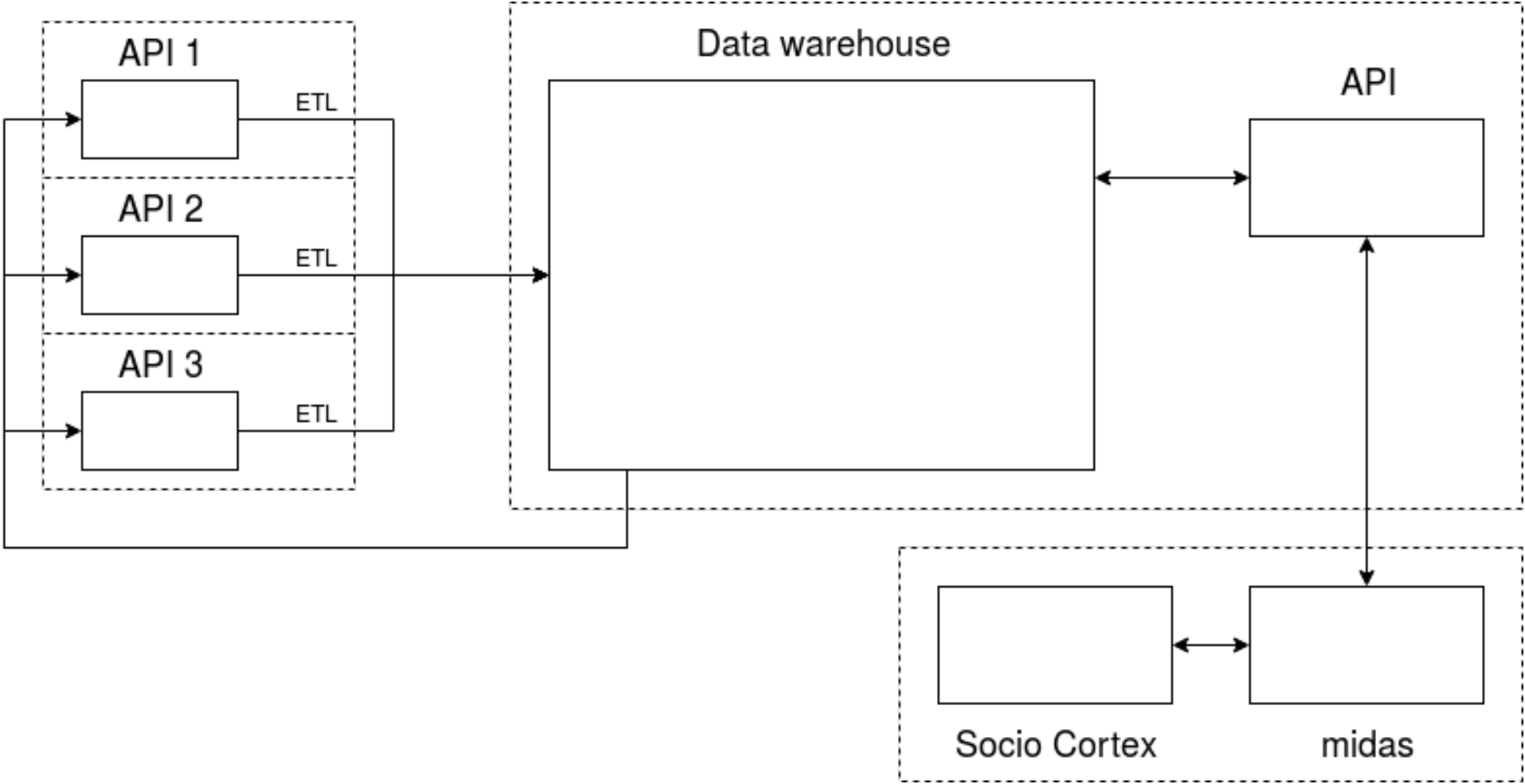


Name	Company	FB	XI	LI	TW	CB	AL	FS
Alex ***	***	X		X	X	X	X	X
Benjamin ***	***			X				
Bernd ***	***	X	X	X	X	X	X	X
Boris ***	***	X	X	X				
Christian ***	***		X	X	X			
Christoph ***	***		X	X	X			
Christoph ***	***		X	X				
Clemens ***	***		X					
Daniel ***	***		X	X			X	X
Daniel ***								
Dominik ***		X		X	X	X	X	X
Florian ***				X	X		X	X
Florian ***	***	X	X	X	X			
Fritz ***	***	X		X	X			
Joachim ***	***		X	X				
Julia ***	***			X				
Kosta ***	***							
Marco ***	***	X	X	X	X	X	X	X
Marco ***	***		X		X			
Maximilian ***	***	X	X	X	X	X	X	
Omri ***	***	X		X	X			
Paul ***	***		X	X	X	X		
Philipp ***	***				X	X		
Roman ***	***		X	X	X			
Sebastian ***	***		X	X				
Stefan ***	***	X						
Vitus ***	***	X	X	X	X	X	X	X
Total ¹		10	16	21	16	8	8	7

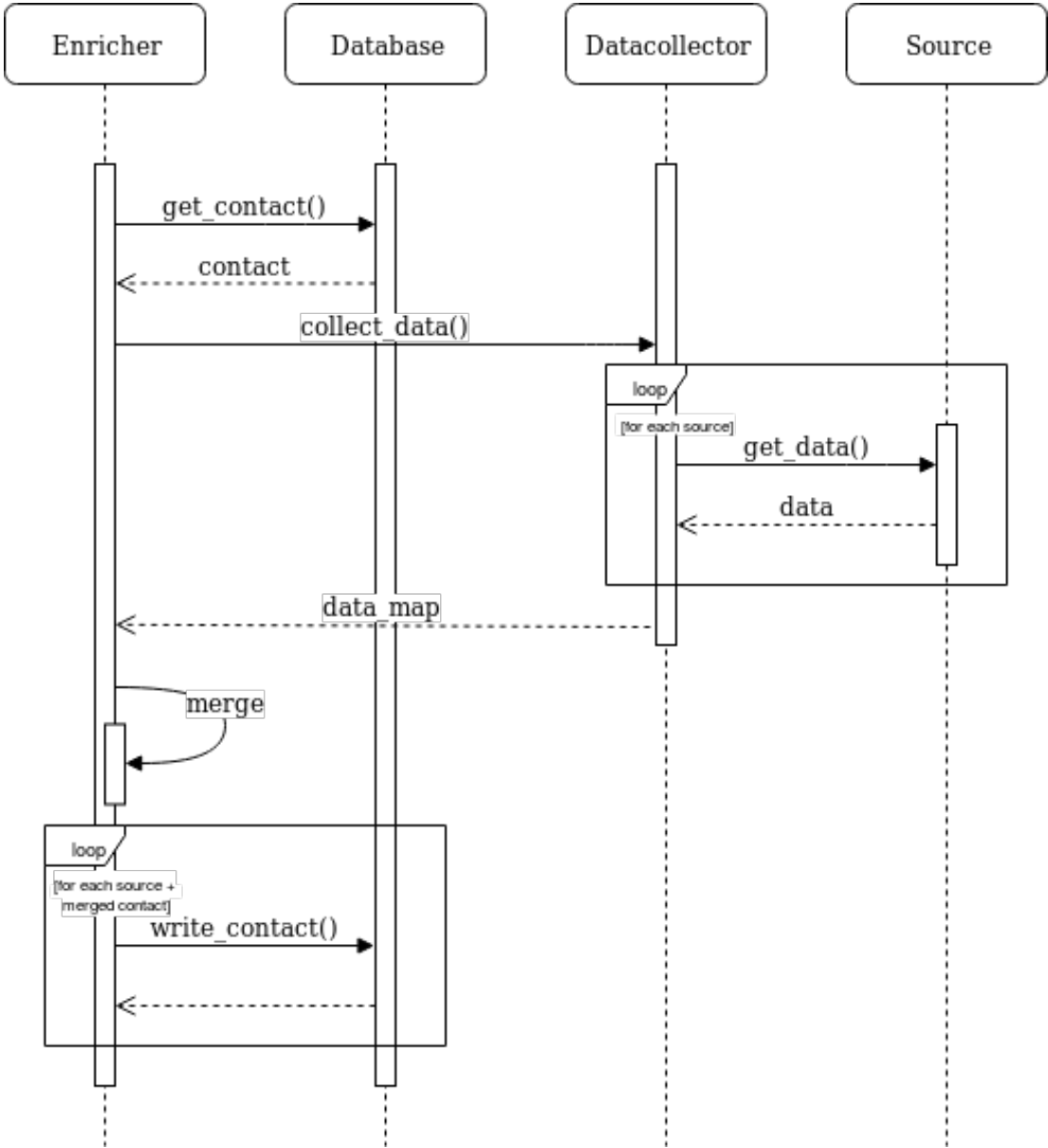
Sources API analysis

- Sources:
 - LinkedIn **X**
 - Xing **X**
 - Twitter ✓
 - Facebook ~
 - Crunchbase ✓
 - AngleList **X**
 - F6S **X**

Prototype architecture



Sequence



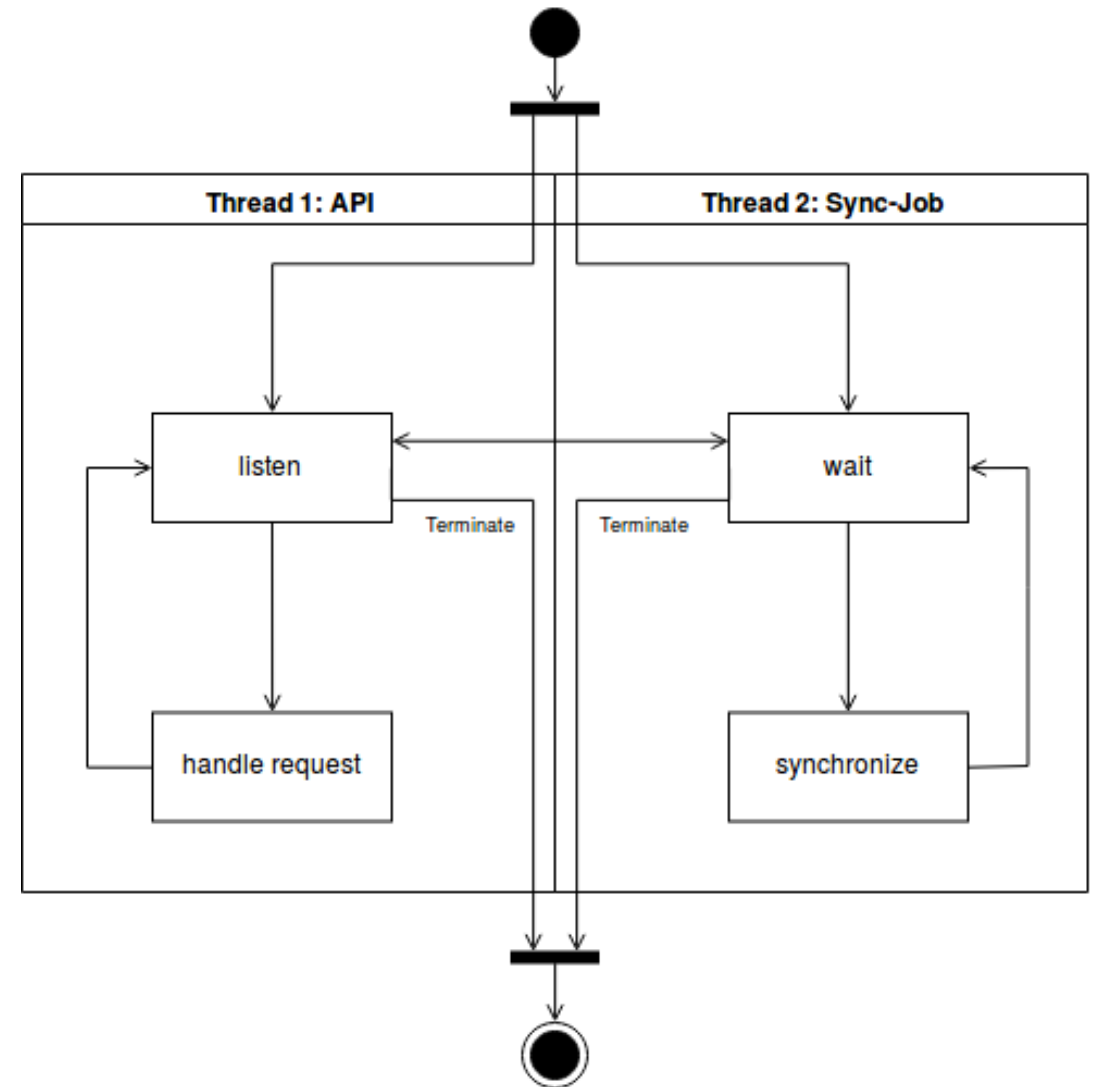
Data quality assessment

- Different merge strategies
- Weighting of data
- ‚Not null‘ metric

Synchronization approach

- Sync interval
- Updating the interval
- $\text{next_sync} = \text{last_sync} + \text{sync_interval}$

contact_id	next_sync
12345678	1523059200
87654321	1523060789
...	...



**Thanks for the
attention**

- <https://www.smartinsights.com/wp-content/uploads/2018/02/DIGITAL-IN-2018-001-GLOBAL-OVERVIEWpng.png>

```
markus@ehrmPC ~/MEGA/TUM/BA/Software/gordias (git)-[master] % curl --data "contacts={
\"contacts\" : [{\"first_name\" : \"Markus\", \"last_name\" : \"Ehringer\", \"organizat
ion\" : \"Technical University of Munich\"}]]}" http://localhost:8080/contacts --data
"age=0.0001"
{"data": [{"first_name": "Markus", "last_name": "Ehringer", "keywords": "bitcoin, tra
ding", "twitter_url": "https://twitter.com/YembooMarkus", "profile_image_urls": "http
://abs.twimg.com/sticky/default_profile_images/default_profile_normal.png", "orga_nam
e": "Technical University of Munich"}]]}%
```


Name	Twitter	Crunchbase		TW	CB
Alex ***	X	X			X
Bernd ***	X	X		X	
Christian ***	X				
Christoph ***	X			X	
Dominik ***	X	X		X	
Florian ***	X				
Florian ***	X			X	
Fritz ***	X				
Marco ***	X	X		X	X
Marco ***	X				
Maximilian ***	X	X		X	X
Omri ***	X				
Paul ***	X	X			X
Philipp ***	X	X		X	X
Roman ***	X				
Vitus ***	X	X			
Total	16	8		7	5

Crunchbase:

- First name
- Last name
- Job
- Profile image URL
- Homepage
- City
- Country
- Facebook profile URL
- Twitter profile URL
- LinkedIn profile URL
- Crunchbase profile URL
- Organization
 - Name
 - City
 - Country
 - Crunchbase profile URL
 - Homepage

Twitter:

- First name
- Last name
- Profile image URL
- Twitter profile URL
- Homepage
- City
- Country
- Keywords