Semi-automatic Generation of Data-Intensive APIs

Shumet Tadesse

Supervisors: Prof. Oscar Romero, Prof. Cristina Gomez and Prof. Katja Hose

eBISS 2019 - Berlin, 5th July 2019

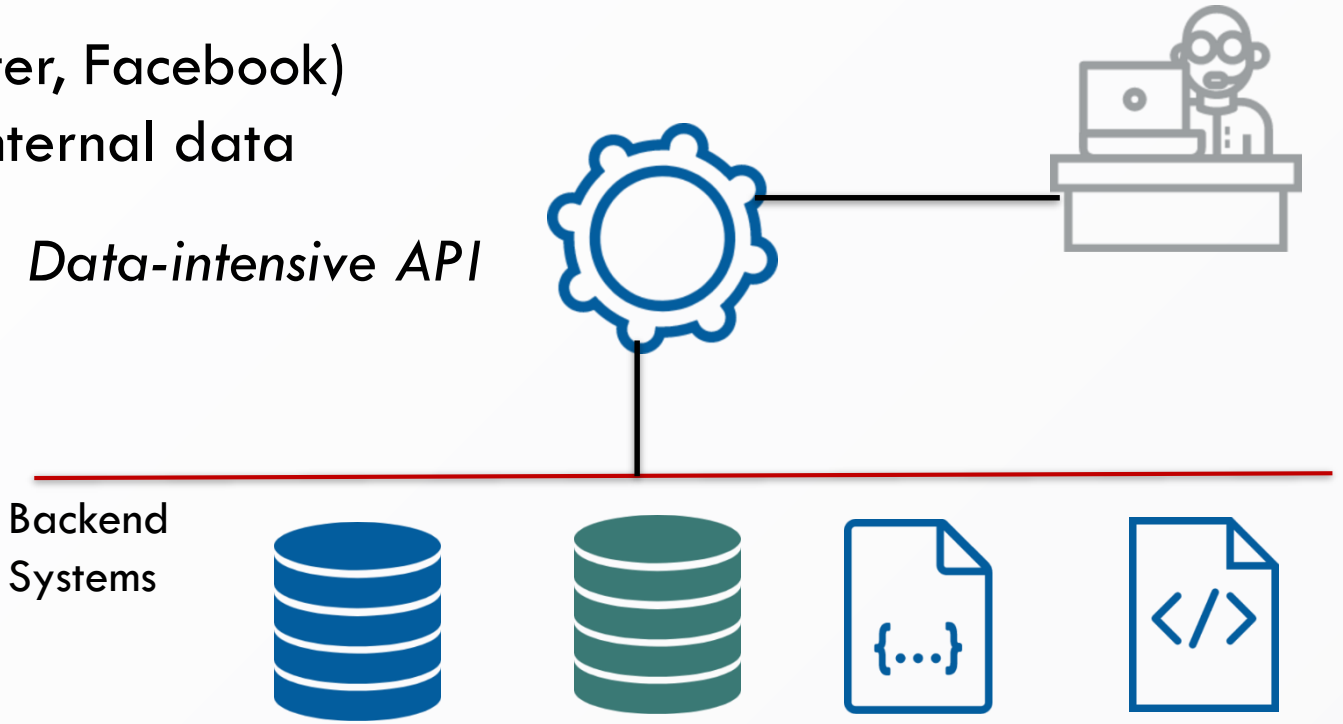
Outline

- Context
  - Data-intensive APIs
  - Challenges
  - Proposed solution
- First Step
  - Background
  - Our approach
  - ARDI Architecture
- Next Steps
- Publications

Context: Data-intensive APIs

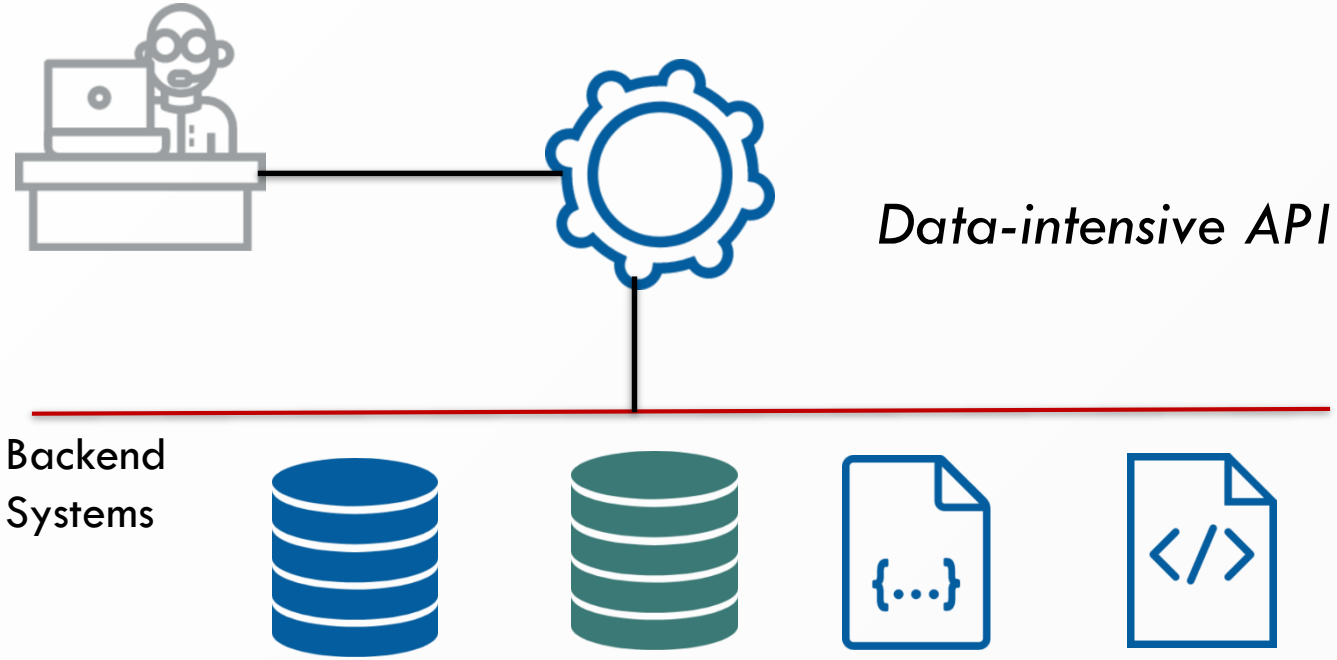
- API is a set of rules, protocols and tools that enable interactions between applications
- At the same time, Businesses build APIs for their customers, or for internal use

- Social Networks (such as Twitter, Facebook) rely on APIs to expose their internal data sources



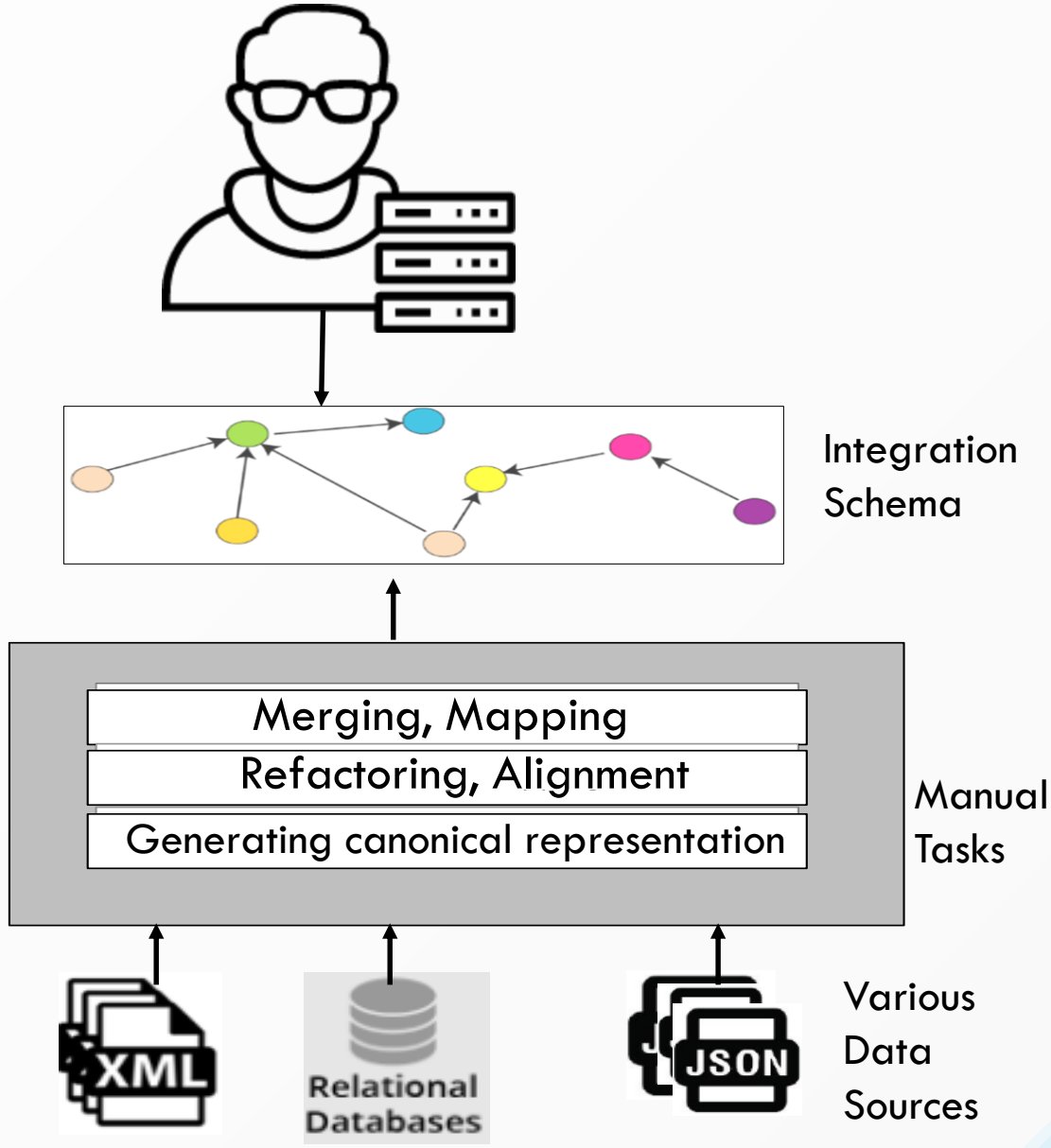
Context: Challenges

- However, building data-intensive APIs is time-consuming and burdensome
- Data-Intensive APIs have traditionally been created manually
- It can be reduced to the Data Integration Problem
  - needs to deal with highly heterogeneous data sources



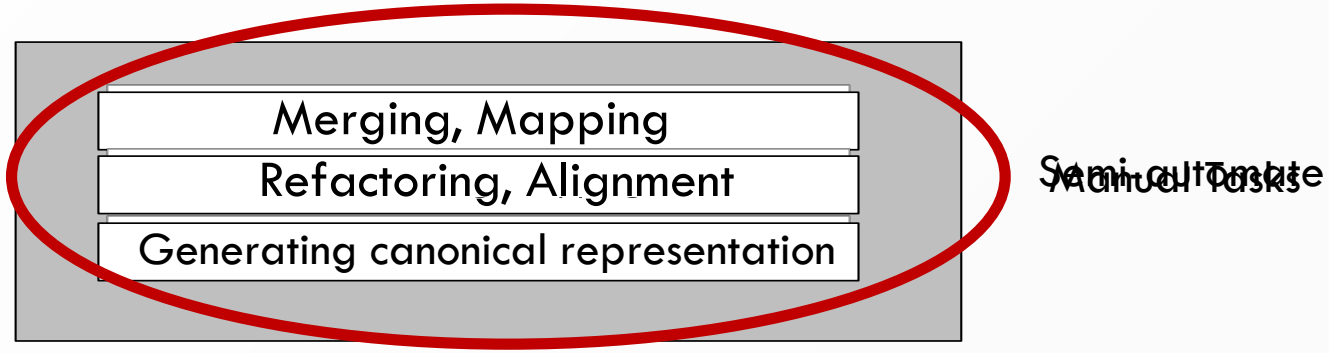
Challenges(2)

- Data Integration is a means to an end
  - expressing each data source in terms of a canonical data model
  - creating a single unified view of the sources, and
  - mapping the data sources to the target schema



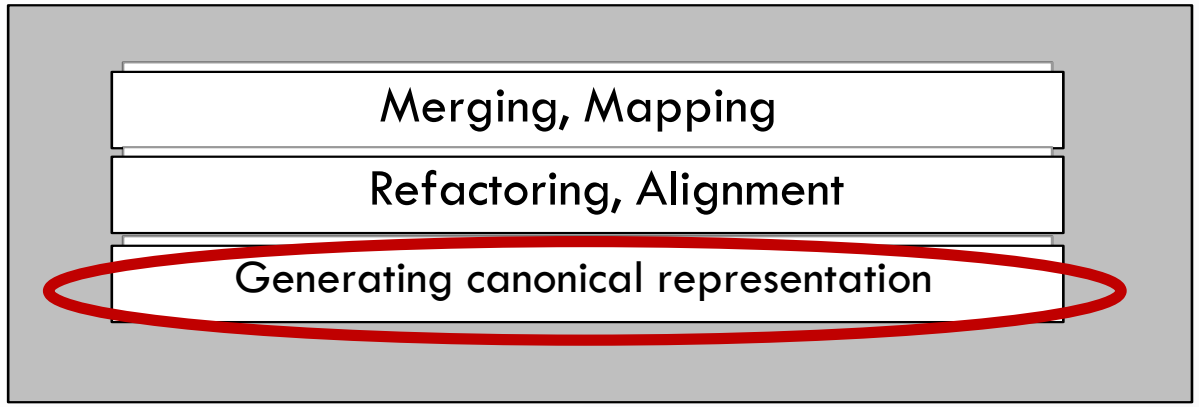
Proposed Solution

- There is a need for systems to automate as much as possible the cumbersome and time-consuming task of integrating heterogeneous data



Golshan, B., Halevy, A., Mihaila, G., Tan, W.C.: Data integration: After the teenage years. In: SIGMOD-SIGACT-SIGAI. pp. 101–106. ACM (2017)

ARDI: Automatic Generation of RDFS Models from Heterogeneous Data Sources for Data Integration



Background

- Source data typically come in terms of schemaless data models such as XML or JSON
  - For schemaless data formats there is typically no available meta-data
- Semantic modeling languages become a key technology for data standardization and conceptualization
- Semantic web community has overlooked the need to generate schema information from data sources automatically

Background(1)

- Approaches for moving data sources to the Semantic Web
  - instance-level: generate a semantic representation of the data (instances)
  - schema-level: translate schema information
- Schema-level approaches, however
  - do not guarantee to produce meta-model compliant schemas,
  - do not fully cover all schema elements that we may find in semi-structured data models (e.g., arrays in JSON), and
  - ignore the RDFS meta-model
- We follow a meta-modeling approach

Our Approach: Why meta-modeling?

- The capability of supporting different abstraction levels
- Helps to maximize the extent to which data can be integrated by separately expressing schema information and the data itself
- Ensures interoperability

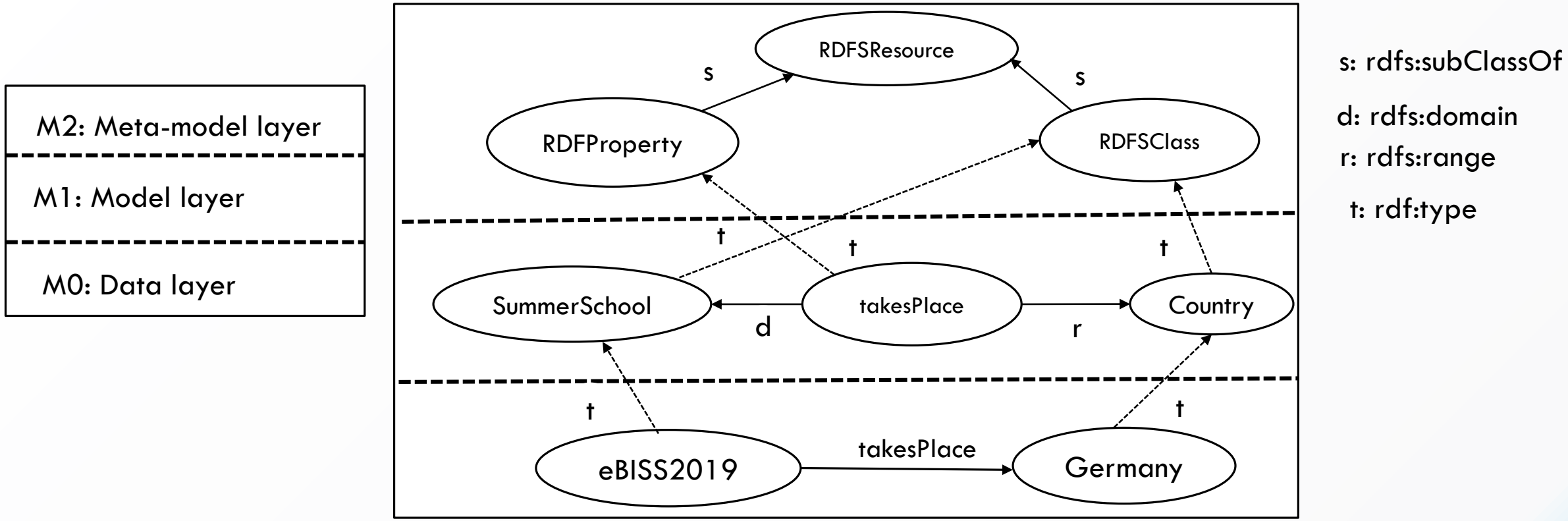
- From a technical point of view:
  - help to minimize development time and
  - maximize efficiency and productivity

Chang, D.T., Kendall, E.: Metamodels for rdf schema and owl. In: MDSW 2004, Monterey, USA (2004)

Our Approach: RDFS as a canonical data model

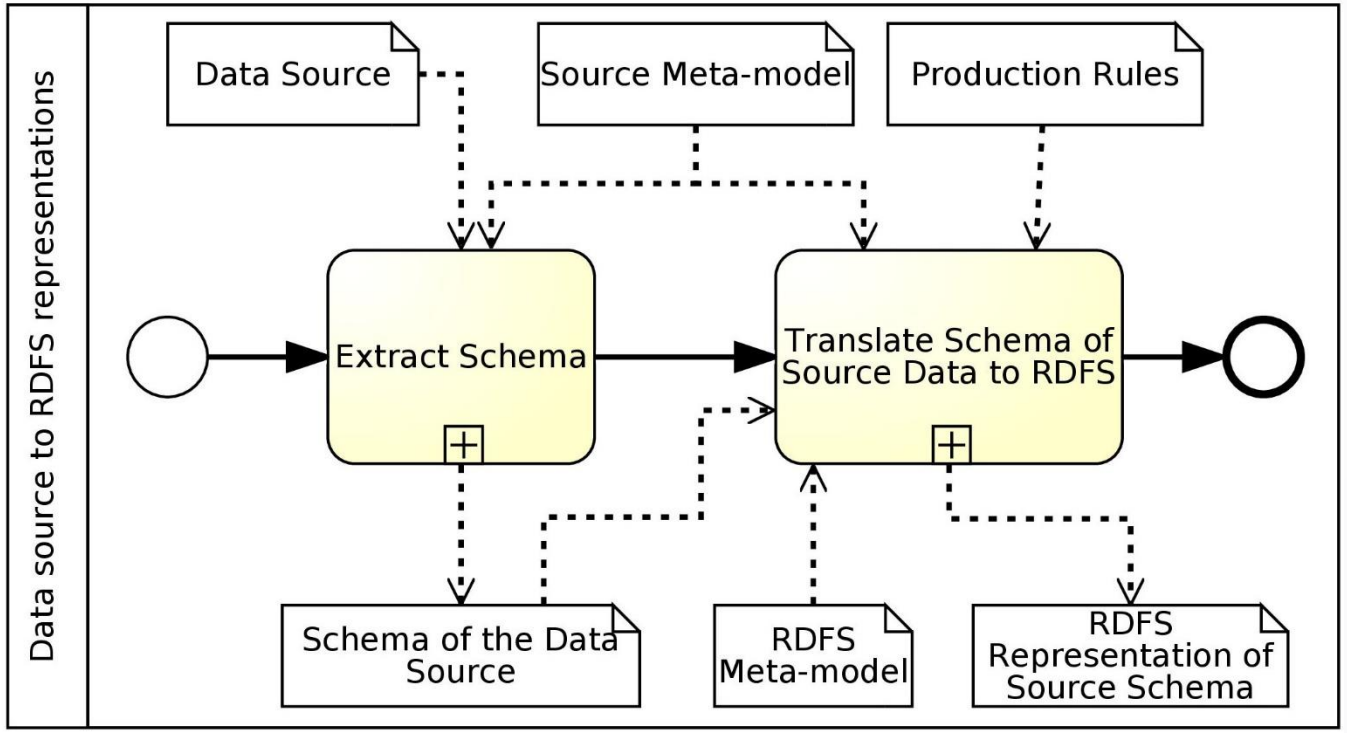
- Expressive
- Flexible
- Non-explicit knowledge can be inferred from explicitly asserted knowledge
- Allows meta-modelling

Our Approach: RDFS Metamodeling



ARDI Workflow

- Extract representations of the sources conformant to the source meta-schema
- Translate to the target schema conformant to the target meta-schema

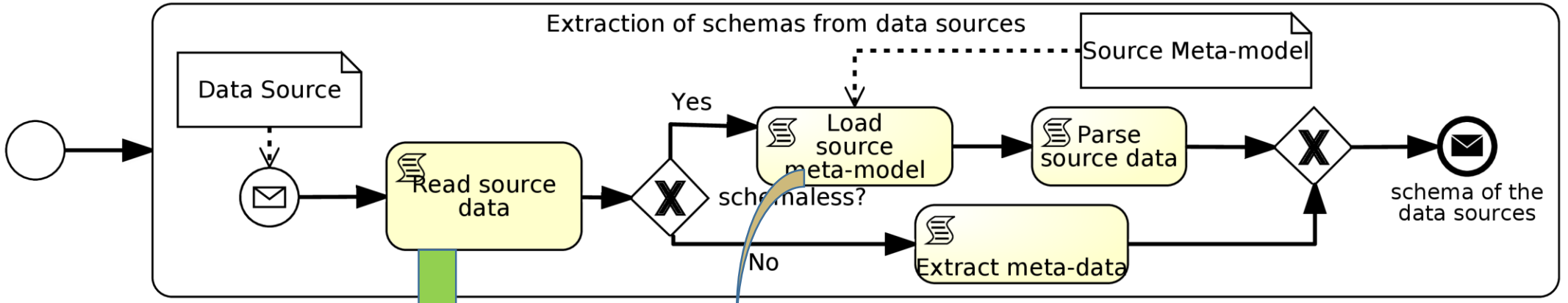


Running Example

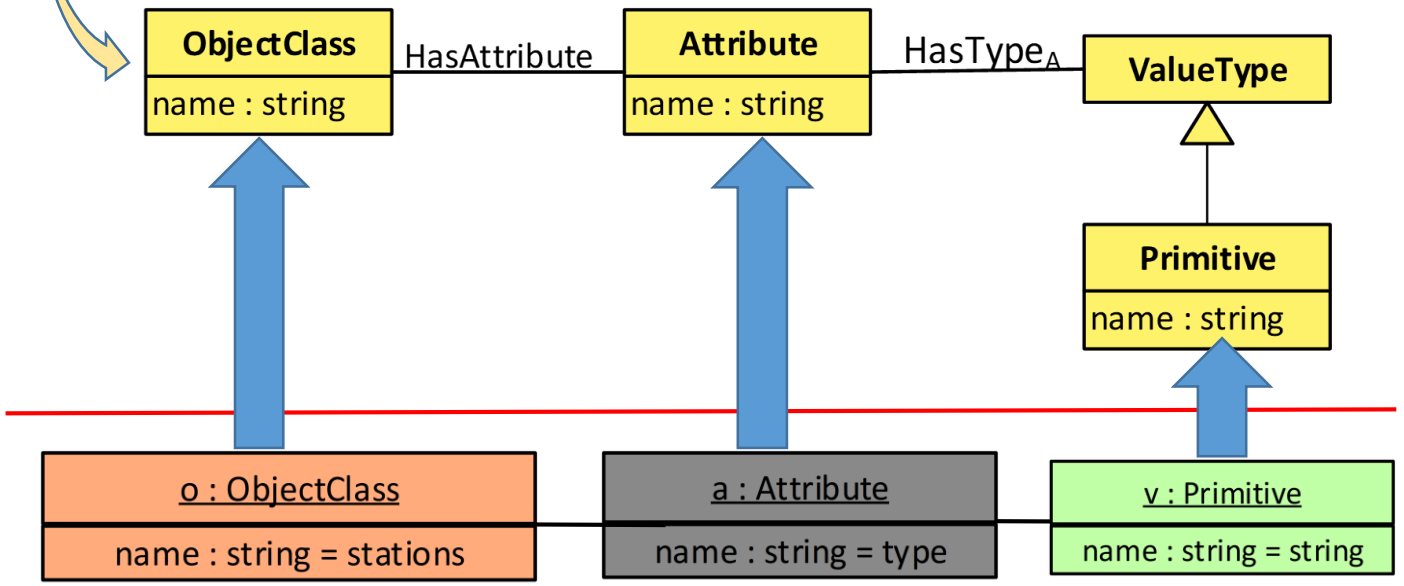
- Stations: attributes with primitive, reference to an object class and array


```
{
   "id":1,
   "type":"BIKE",
   "address":{
      "streetName":"Gran Via Corts Catalanes",
      "streetNumber":760
   },
   "coordinates":[
      41.397952,
      2.180042
   ],
   "nearbyStations":[
      {
         "id":24,
         "type":"Metro",
         "distance":500
      },
      {
         "id":426,
         "type":"Bus",
         "distance":367
      }
   ]
}
```

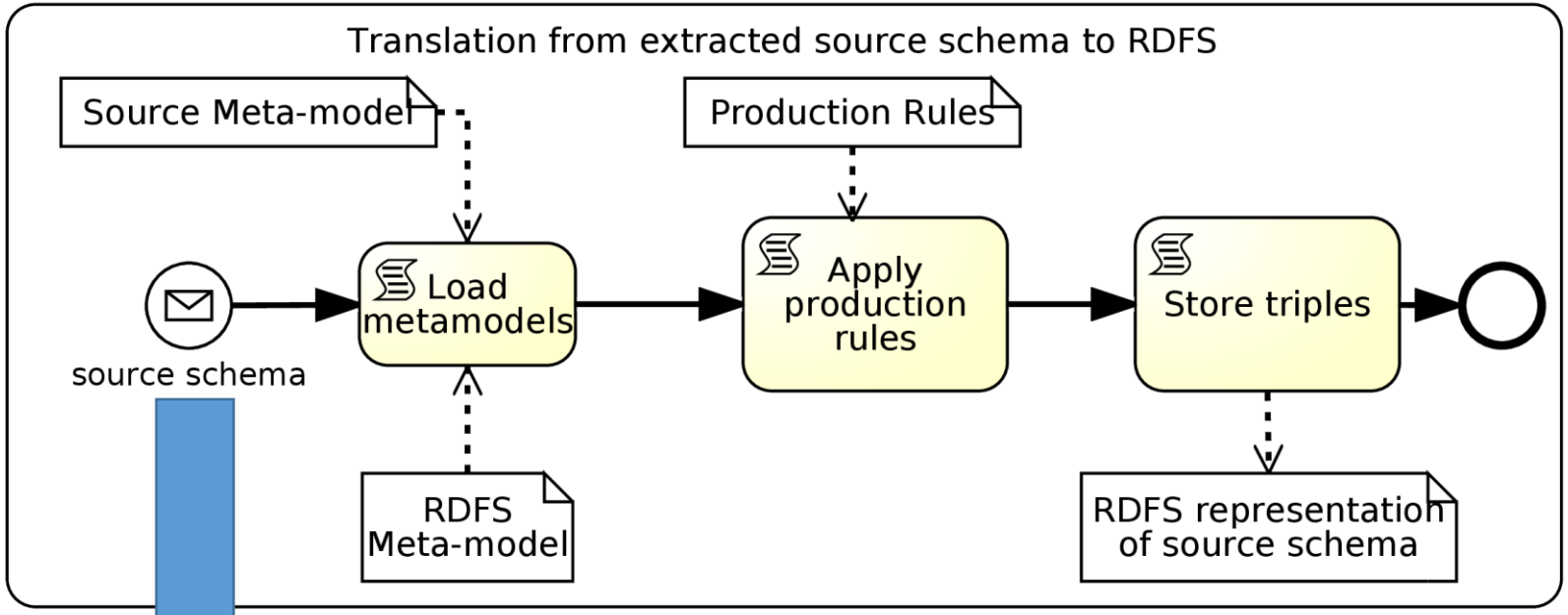
Extraction of Schema

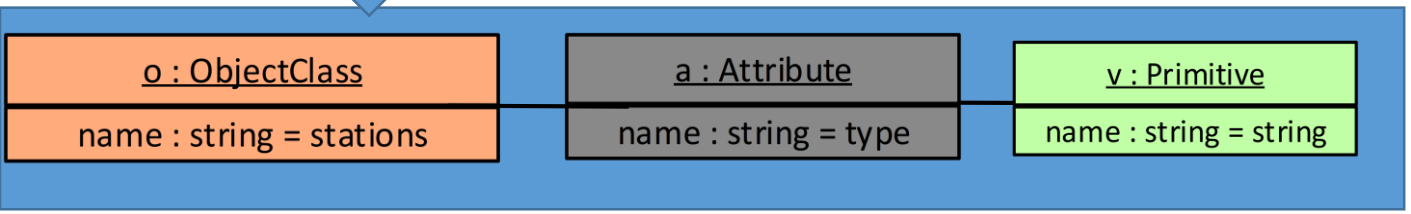



```
{
   "id":1,
   "type":"BIKE",
   "address":{
      "streetName":"Gran Via Corts Catalanes",
      "streetNumber":760
   },
   "coordinates":[
      41.397952,
      2.180042
   ],
   "nearbyStations":[
      {
         "id":24,
         "type":"Metro",
         "distance":500
      },
      {
         "id":426,
         "type":"Bus",
         "distance":367
      }
   ]
}
```

Translation of Schema





R_1

o : ObjectClass
name : string = stations

sc: stations a rdfs:Class .

R_2

a : Attribute
name : string = type

sc:stations/type a rdf:Property ; rdfs:domain sc: stations .

R_3

v : Primitive
name : string = string

sc:stations/type rdfs:range xsd:string

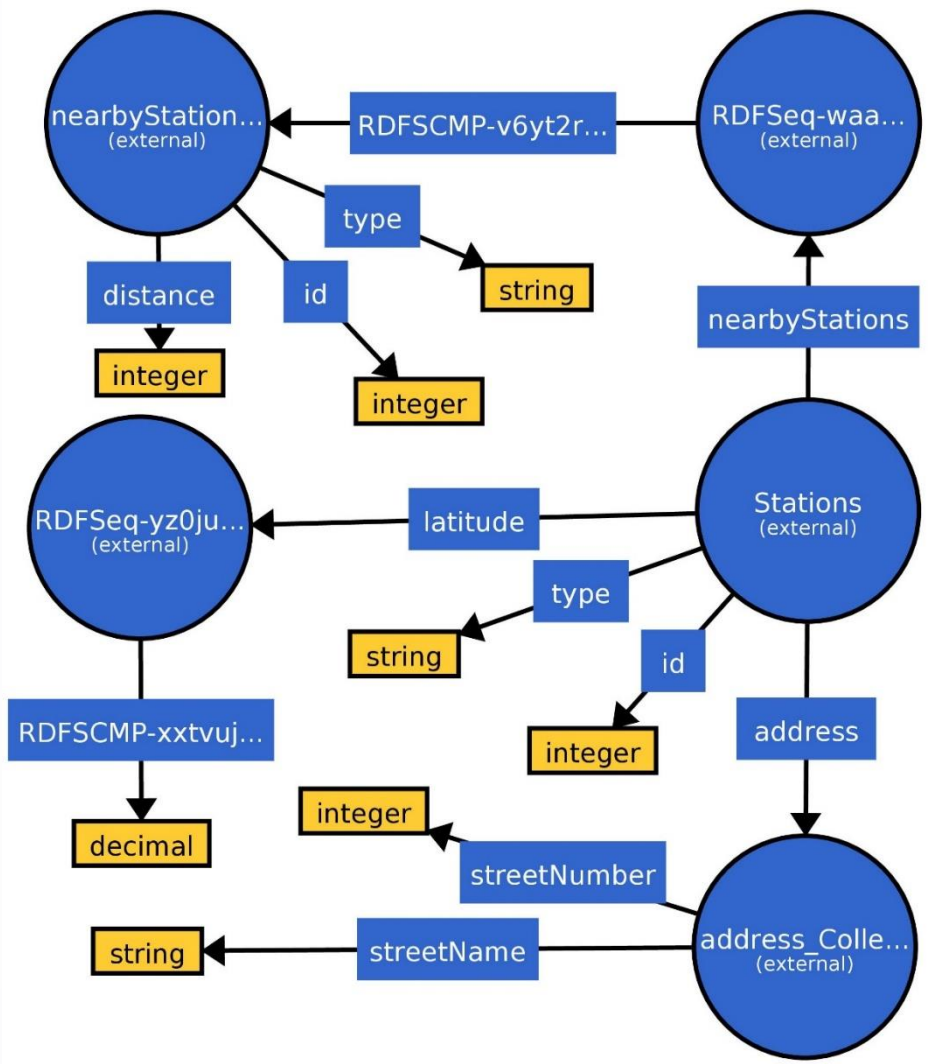
Production Rules

- Define the translation from the schema of the source data to equivalent RDFS representation
- Formalized in First Order Logic
- Represented as a logical axiom with left-hand side(LHS) and right-hand side (RHS)
  - if LHS holds RHS must hold too

Prototype Instantiation

stations.json


```
{
   "id":1,
   "type":"BIKE",
   "address":{
      "streetName":"Gran Via Corts Catalanes",
      "streetNumber":760
   },
   "coordinates":[
      41.397952,
      2.180042
   ],
   "nearbyStations":[
      {
         "id":24,
         "type":"Metro",
         "distance":500
      },
      {
         "id":426,
         "type":"Bus",
         "distance":367
      }
   ]
}
```

RDFSDatatype

RDFS Class

RDFSProperty

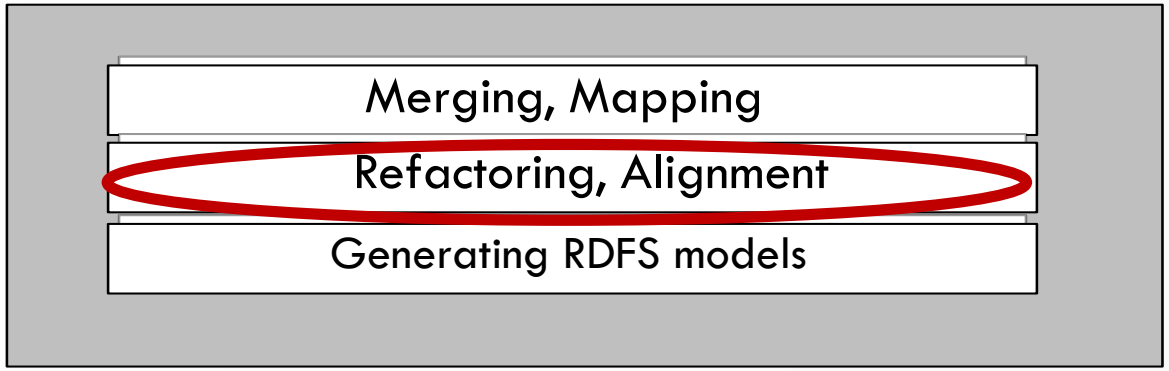
RDFSRange

RDFSDomain

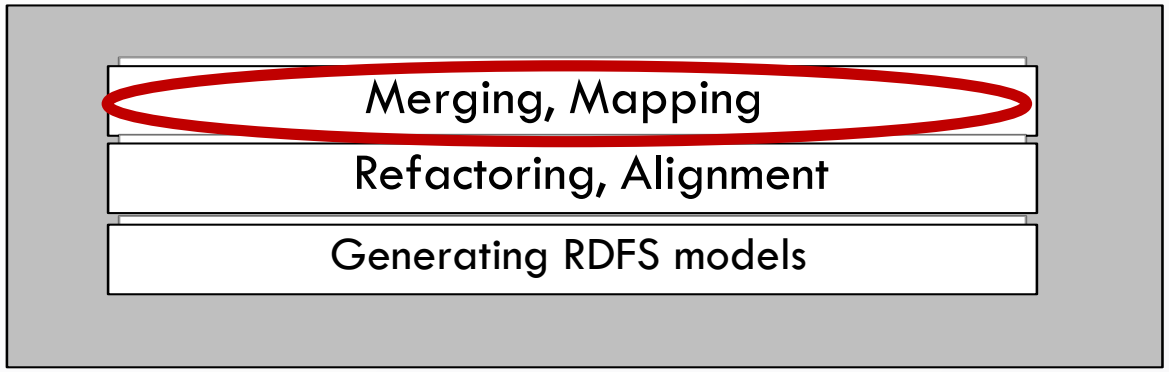
Next Steps

- Refactoring automatically extracted source representations
  - resemble the physical structure of the underlying data sources
  - a richer representation of domain concepts and relationships is required to integrate lately

- Alignment



- Integrating and querying the source representations
  - Merging
  - Mapping



Publications

Submitted:

Shumet Tadesse, Cristina Gomez, Oscar Romero, Katja Hose "**ARDI: Automatic Generation of RDFS Models from Heterogeneous Data Sources**" IEEE EDOC 2019

Planned:

Conference Paper II: **Enhancing Data Integration by Refactoring Automatically Extracted Ontologies**

- Authors: Shumet Tadesse, Cristina Gomez, Oscar Romero, Katja Hose
- Outlet: The International Conference on Extending Database Technology (EDBT), October 2019

Journal Paper: **Automatically Generating data-intensive APIs**

- Authors: Shumet Tadesse, Cristina Gomez, Oscar Romero, Katja Hose
- Outlet: Journal of Systems and Software (JSS), December 2019

Conference Paper III: **Supporting the Automation of the Whole Data Integration Life-Cycle**

- Authors: Shumet Tadesse, Cristina Gomez, Oscar Romero, Katja Hose
- Outlet: The International Semantic Web Conference (ISWC), April 2020

Demo Paper: **Integrating Heterogeneous Data Sources for the Generation of data-intensive APIs**

- Authors: Shumet Tadesse Nigatu, Cristina Gomez, Oscar Romero, Katja Hose
- Outlet: Conference on Advanced Information Systems Engineering (CAiSE), November 2020

Thank You!