Introduction to
Data Integration

Mustafa Jarrar

Birzeit University

Watch this lecture and download the slides

Online Courses : http://www.jarrar.info/courses

Thanks to Anton Deik for helping me preparing this lecture

Data Integration

Part 1: Example of a Data Integration Problem

Part 2: Challenges of Data Integration:

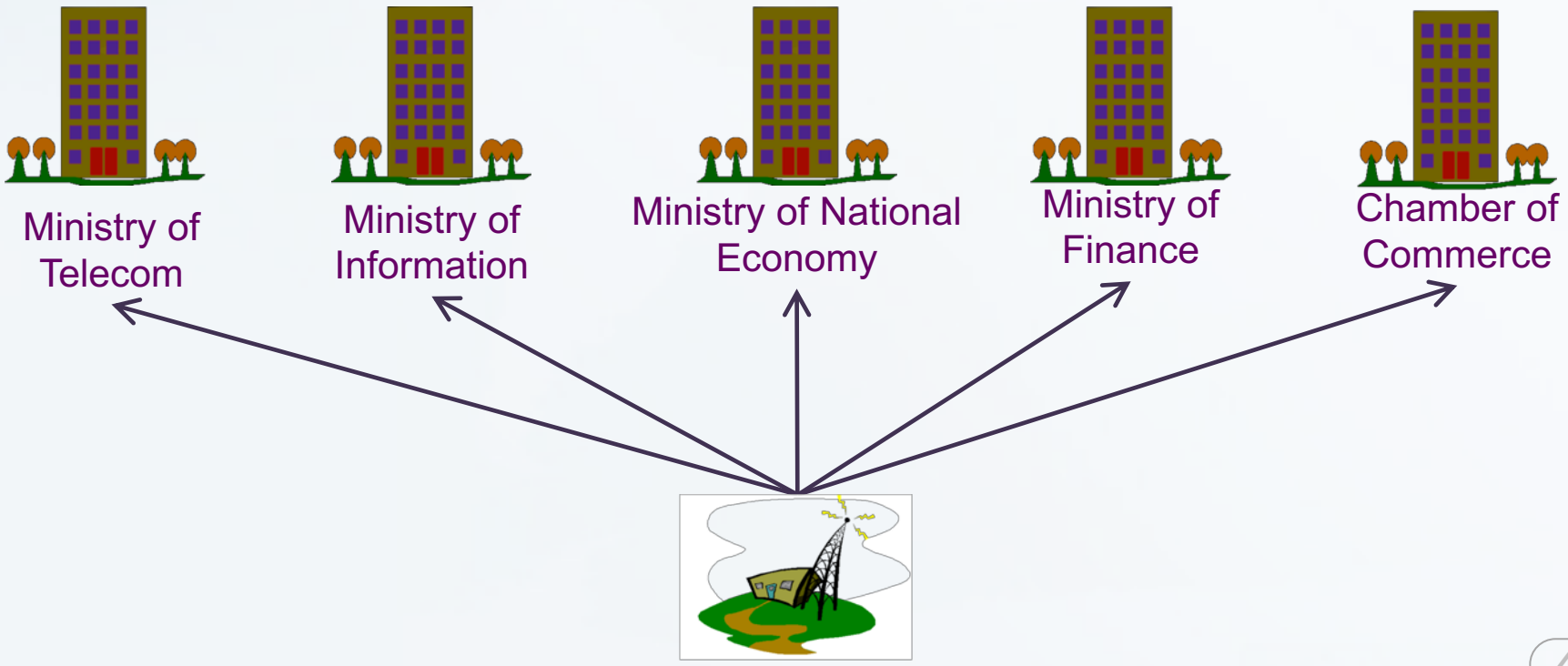
- Heterogeneities in Database Schemas
- Heterogeneities in Name and Meaning
- Heterogeneities in Structure and Type
- Heterogeneities in the rules and constraints
- Heterogeneities in data model

Keywords: Data Integration, Registered data, domain, domain name system, web, distributed database, database schema, Heterogeneities, Model Heterogeneities, Data model, Synonyms, Homonyms, Attribute, Entity

Example from the government Domain

Consider all interactions with government agencies in order to register a new business.

Example: Establishing a new Radio Station in Palestine, which involves 5 agencies.

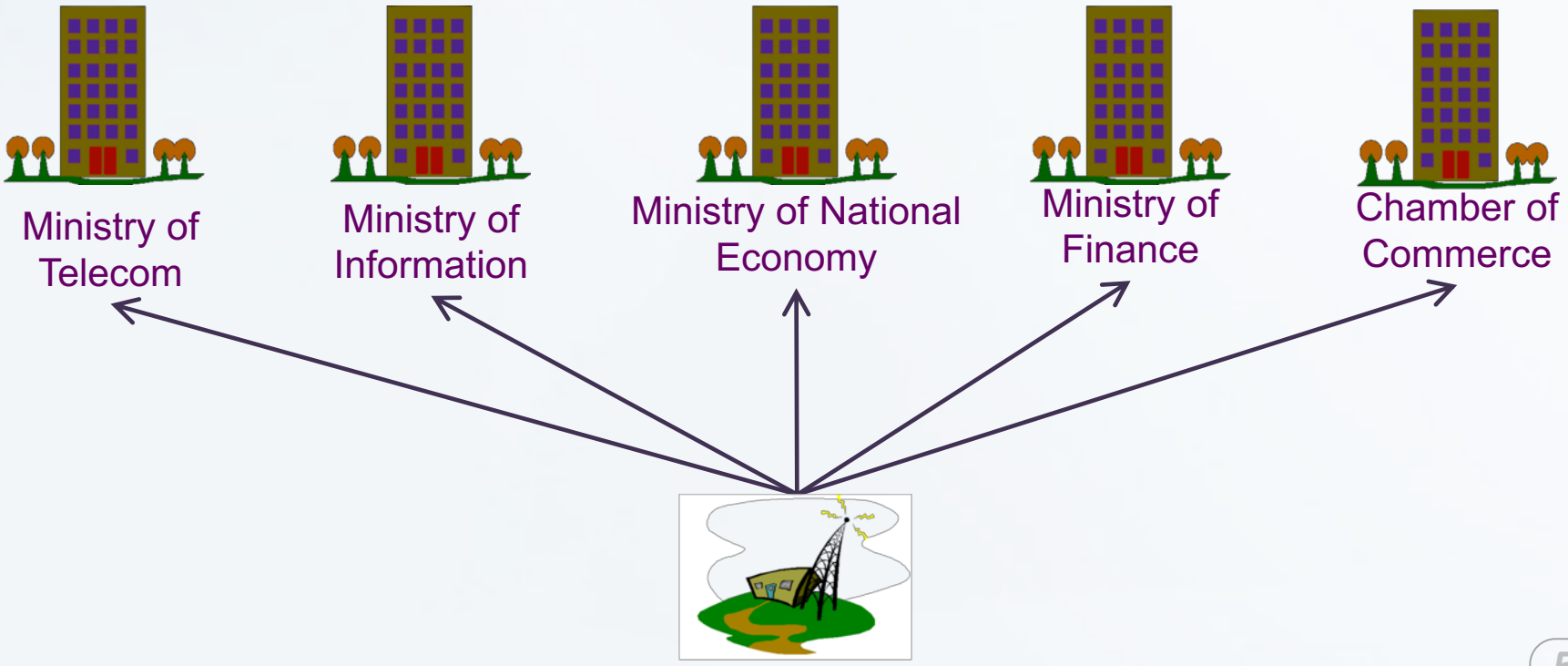


Example from the government Domain

When your business evolves or changes.

Example: Changing the address of the radio station.

- Address must be changed in 5 different databases.



Example from the government Domain

Consider the data registered about the same radio station in the databases of different ministries and governmental agencies:

ID	Name	Type	City
R2563I	Radio Al-Amal	Radio Station	Ramallah

B_ID	Business Name	Activity Type	Province
LM1847	Al-Amal Broadcast	Radio Broadcasting	Ramallah and Bireh

ID	Company Name	Company Type	Location
182NS3	Broadcast Al-Amal	Broadcasting Station	Al-Balu’

. . .

Example from the government Domain

From our simple example one can point out to some challenges in Data Integration:

- No agreed upon naming (name, business name, company name)
- No agreed upon meaning (Does 'Activity Type' mean exactly the same as 'Company Type'?)
- Different Registered Data: Radio Al-Amal, Al-Amal Broadcast, ….

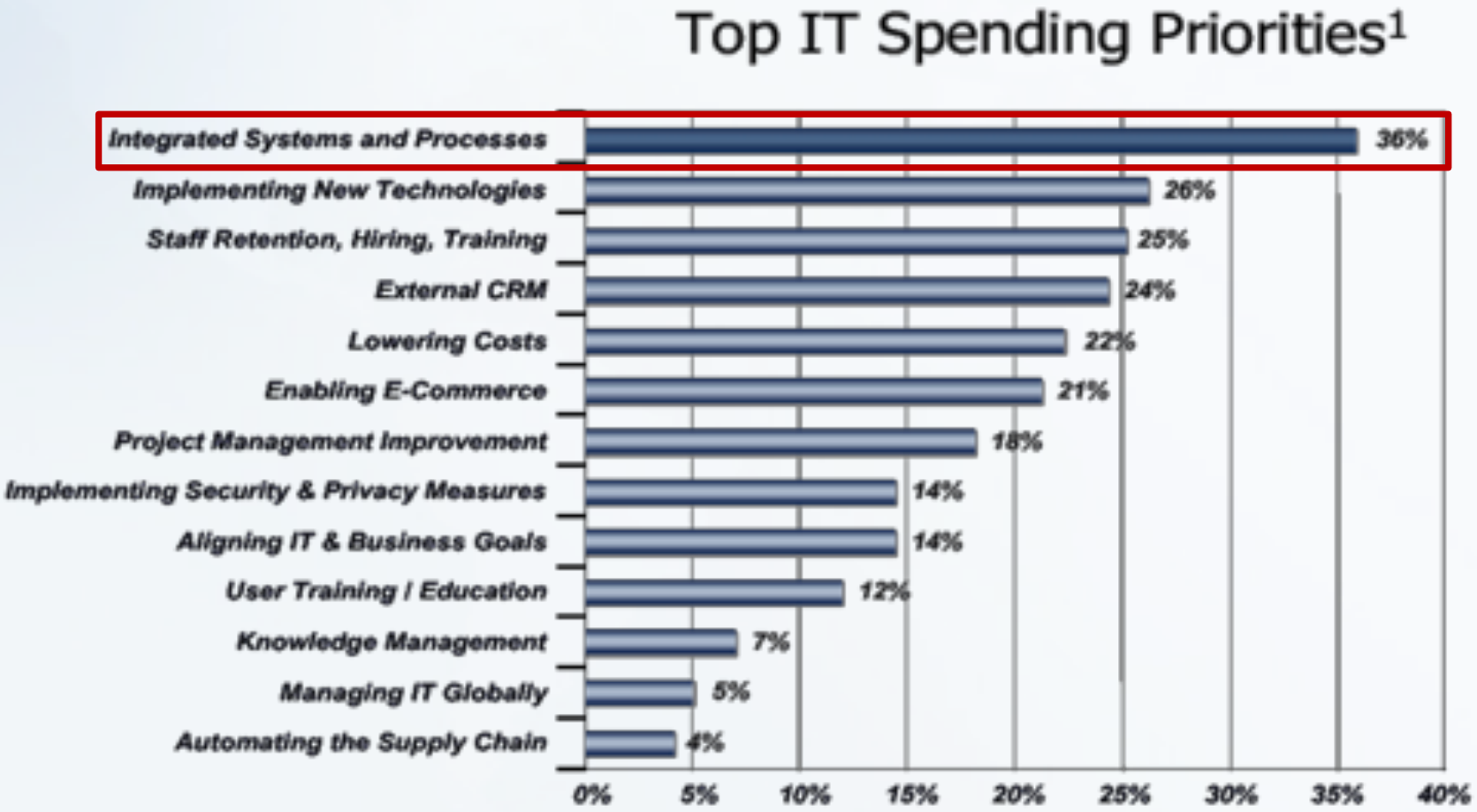
ID	Name	Type	City
R2563I	Radio Al-Amal	Radio Station	Ramallah

B_ID	Business Name	Activity Type	Province
LM1847	Al-Amal Broadcast	Radio Broadcasting	Ramallah and Bireh

ID	Company Name	Company Type	Location
182NS3	Broadcast Al-Amal	Broadcasting Station	Al-Balu'

• • •

Problem is in all domains



[1]CIO Magazine Survey, February 2002

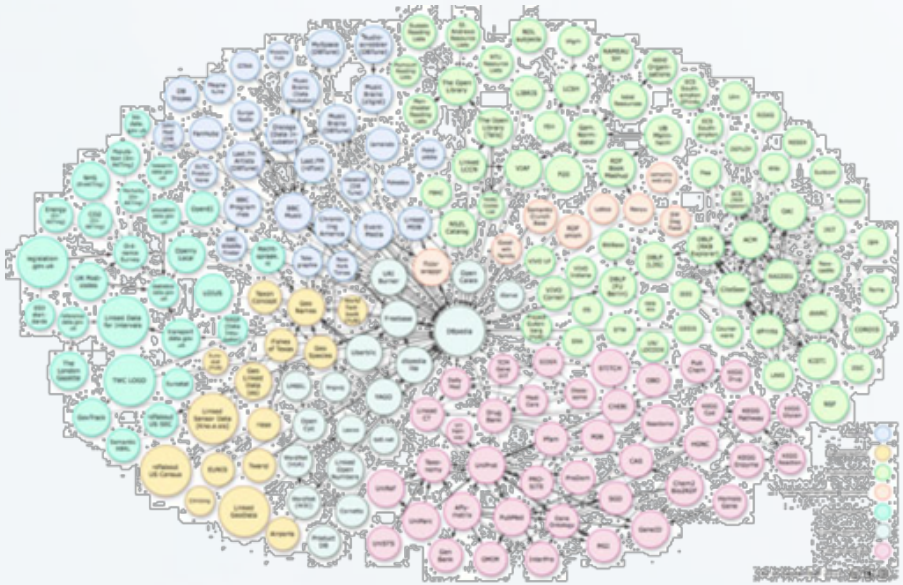
33% of firms surveyed have EAI projects (Forrester, March 2002 Business Technographics benchmark)

Problem is in all domains

Problem is now even more challenging with the Web.

The Data Web envisions the web as a global world-wide database.

This means that one can query distributed multiple databases on the web as if he/she is querying a local database.



Data Integration

Part 1: Example of a Data Integration Problem

Part 2: Challenges of Data Integration:

- Heterogeneities in Database Schemas
- Heterogeneities in Name and Meaning
- Heterogeneities in Structure and Type
- Heterogeneities in the rules and constraints
- Heterogeneities in data model

Keywords: Data Integration, Registered data, domain, domain name system, web, distributed database, database schema, Heterogeneities, Model Heterogeneities, Data model, Synonyms, Homonyms, Attribute, Entity

Challenges of Data Integration:
Heterogeneities in Database Schemas

One can distinguish between several heterogeneities between different schemas:

- Name Heterogeneities (difference in used vocabulary).
- Meaning Heterogeneities (different meaning/semantics for the same attribute in two schemas).
- Heterogeneities in the structure and type.
- Heterogeneities in the rules and constraints.
- Data Model Heterogeneities.

Name and Meaning Heterogeneities

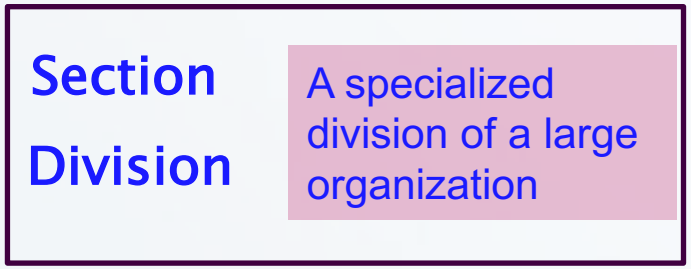
Synonyms – different names for the same concepts

- employee, clerk
- exam, course
- code, num

Homonyms – same name for different concepts (different meanings)

- City as *City of birth* in one schema,
- City as *City of Residence* in another schema

Homonyms



Synonyms

Heterogeneities in Structure and Type

Source: Carlo Batini

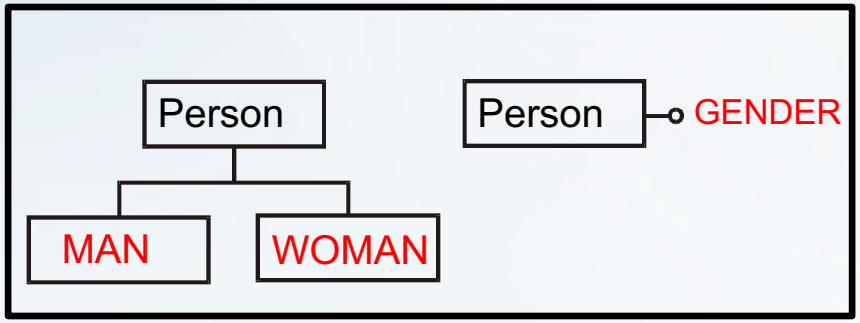
The same concepts are represented with different conceptual structures in two schemas:

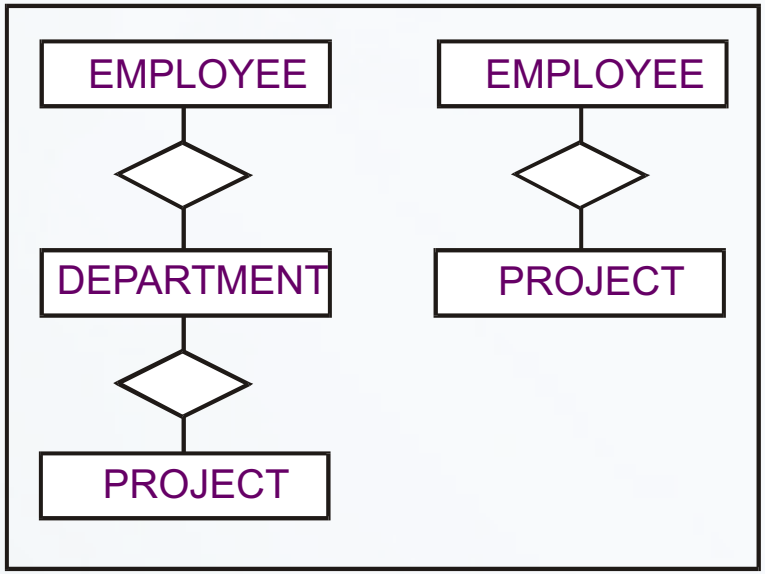
- Attribute in one schema and derived value in another schema.
- Attribute in one schema and entity in another schema.
- Entity in one schema and relationship in another schema.
- Different abstraction levels for the same concept in two schemas: e.g. two entities with homonym names related by an IS-A hierarchy in two schemas.

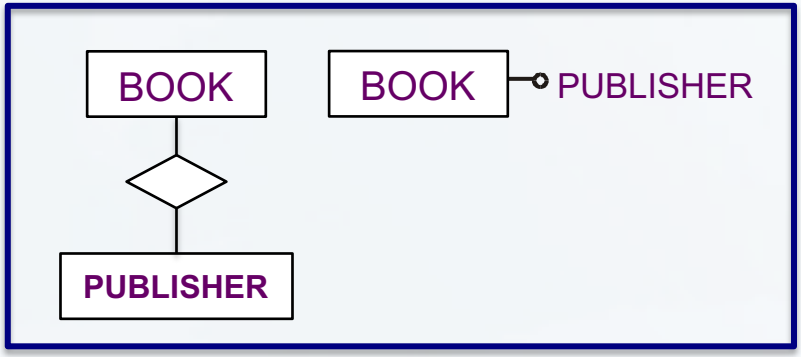
Heterogeneities in Structure and Type

Source: Carlo Batini

EXAMPLES:







Heterogeneities in Type

Examples:

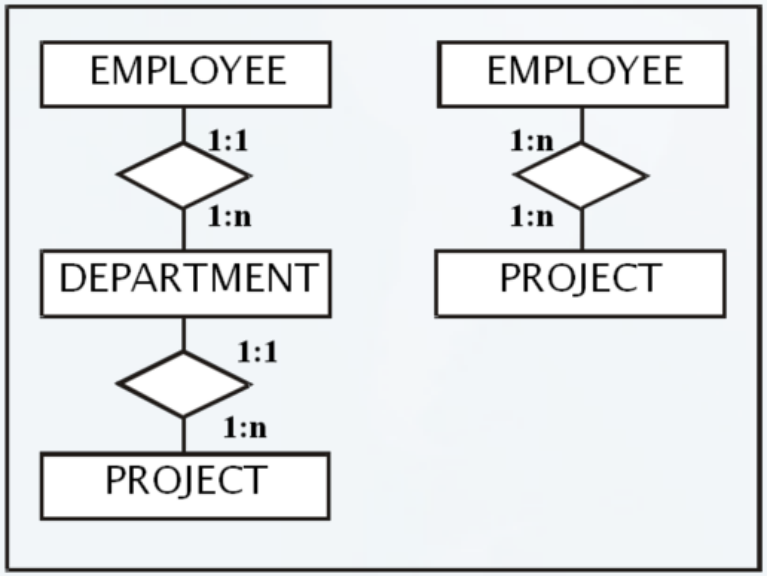
- In a single attribute (e.g., Numberic, Alphanumeric). E.g., the attribute “gender”:
  - Male/Female
  - M/F
  - 0/1
- Year has a four digit domain in one schema and two digit domain in another schema
- Different currencies (Euros, US Dollars, etc.)
- Different measure systems (kilos vs. pounds, centigrade vs. Fahrenheit.)
- Different granularities (grams, kilos, etc.)

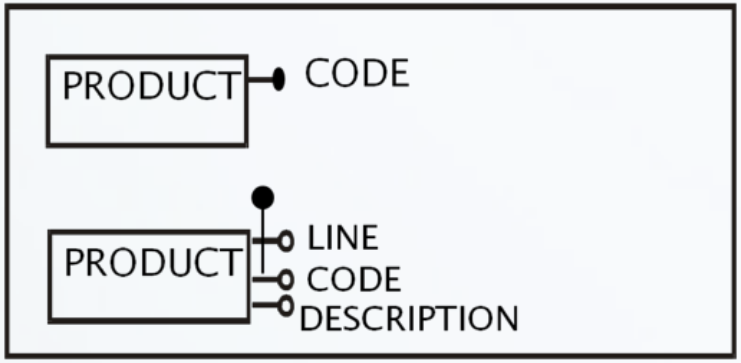
Heterogeneities in the rules and constraints

Source: Carlo Batini

EXAMPLES:

- Different cardinalities in the same relationships
- Key conflicts





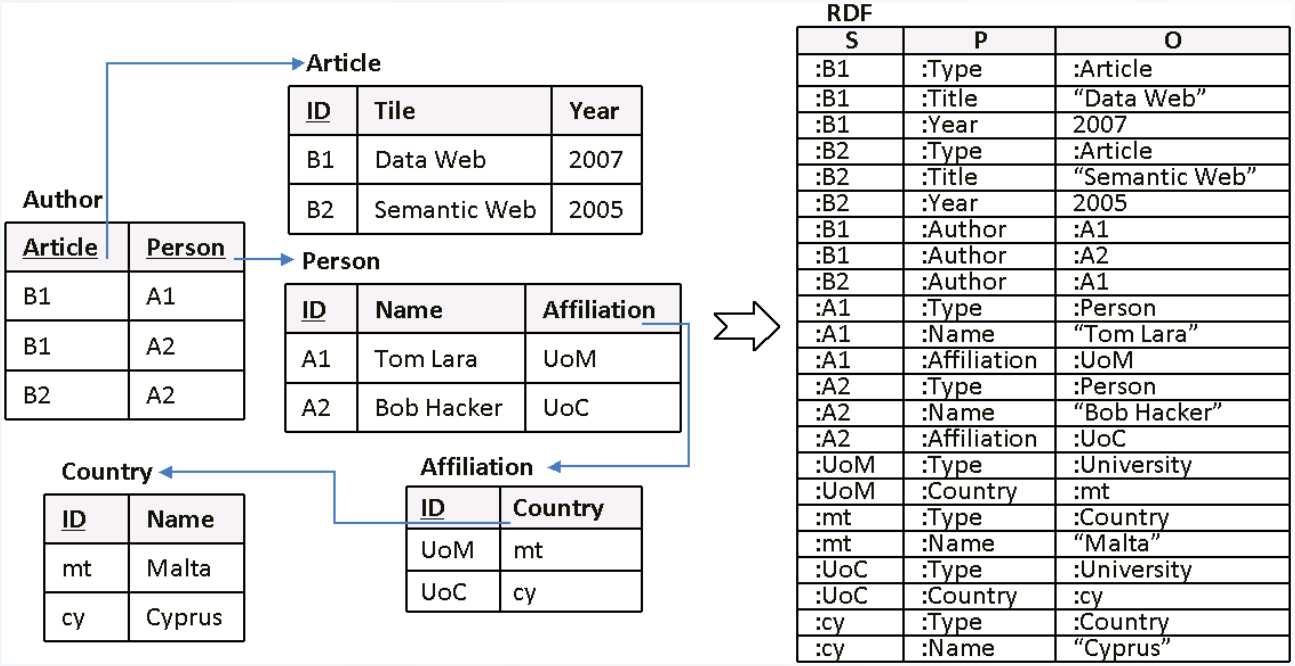
Model Heterogeneities

Model Heterogeneities occurs when different databases adheres to different data models:

- Relational Data Model, XML, RDF, Object-Oriented, OWL, ...

Solution: Reduce Model Heterogeneity by using one data model.

Example: Convert the Relational Model to RDF graph model.



References

[1] Mustafa Jarrar, Anton Deik: The Graph Signature: A Scalable Query Optimization Index for RDF Graph Databases Using Bisimulation and Trace Equivalence Summarization. International Journal on Semantic Web and Information Systems, 11(2), 36-65,. April-June 2015

[2] Mustafa Jarrar, Anton Deik, Bilal Faraj: Ontology-Based Data And Process Governance Framework -The Case Of E-Government Interoperability In Palestine . In pre-proceedings of the IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA'11). Pages(83-98). 2011.

[3] Mustafa Jarrar and Marios D. Dikaiakos: A Query Formulation Language for the Data Web. The IEEE Transactions on Knowledge and Data Engineering. IEEE Computer Society. Pages(783-798). Volume 24, Number 4, April 2012

[4] Paolo Ceravolo, Chengfei Liu, Mustafa Jarrar, Kai-Uwe Sattler: Special Issue on Querying the Data Web -Novel techniques for querying structured data on the web. The World Wide Web Journal. Volume(14), Issue (5-6). Springer. August 2011. ISSN:1573-1413.

[5] Anton Deik, Bilal Faraj, Ala Hawash, Mustafa Jarrar: Towards Query Optimization for the Data Web - Two Disk-Based algorithms: Trace Equivalence and Bisimilarity. Proceedings of the 3rd Palestinian International Conference on Computer and Information Technology (PICCIT 2010). 2010.

[6] Mustafa Jarrar, Marios D. Dikaiakos: Querying the Data Web: the MashQL Approach. IEEE Internet Computing. Volume 14, No. 3. Pages (58-670). IEEE Computer Society, ISSN 1089-7801. May 2010.

[7] Mustafa Jarrar, Marios D. Dikaiakos: Querying the Data Web: the MashQL Approach. IEEE Internet Computing. Volume 14, No. 3. Pages (58-670). IEEE Computer Society, ISSN 1089-7801. May 2010.Mustafa Jarrar and Marios D. Dikaiakos: A Data Mashup Language for the Data Web . Proceedings of LDOW, WWW'09. ACM. ISSN 1613-0073. (2009).

[8] Mustafa Jarrar and Marios D. Dikaiakos: MashQL: a query-by-diagram topping SPARQL -Towards Semantic Data Mashups. Proceedings of ONISW'08, part of the ACM CiKM conference. ACM. pages (89-96) ISBN 9781605582559.(2008).

[0] Mustafa Jarrar: Towards methodological principles for ontology engineering. PhD Thesis. Vrije Universiteit Brussel. (May 2005)

[10] Mustafa Jarrar, Luk Vervenne, Diana Maynard: HR-Semantics Roadmap- The Semantic challenges and opportunities in the Human Resources domain . Technical Report. The Ontology Outreach Advisory, Belgium. (OOA-HR/2007-08-20/v025). August 2007

[11] Lyndon Nixon, Malgorzata Mochol, Mustafa Jarrar, Stamatia Dasiopoulou, Vasileios Papastathis, and Yiannis Kompatsiaris: Prototypical business use cases. Deliverable D1.1.2 (WP1.1), The Knowledge Web Network of Excellence (NoE) IST-2004-507482, Luxemburg. January 2005.

[12] Peter Spyns, Daniel Oberle, Raphael Volz, Jijuan Zheng, Mustafa Jarrar, York Sure, Rudi Studer, and Robert Meersman: OntoWeb- a Semantic Web Community Portal. Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management (PAKM 2002). Pages (189-200). LNCS 2569, Springer. ISBN: 3540003142. December 2002.

[13] Carlo Batini: Course on Data Integration. BZU IT Summer School 2011.

[14] Stefano Spaccapietra: Information Integration. Presentation at the IFIP Academy. Porto Alegre. 2005.

[15] Chris Bizer: The Emerging Web of Linked Data. Presentation at SRI International, Artificial Intelligence Center. Menlo Park, USA. 2009.