

Armazéns de Dados

Departamento de Engenharia Informática (DEI/ISEP)
Paulo Oliveira
pjo@isep.ipp.pt

1

Taxonomy of Data Quality Problems

2

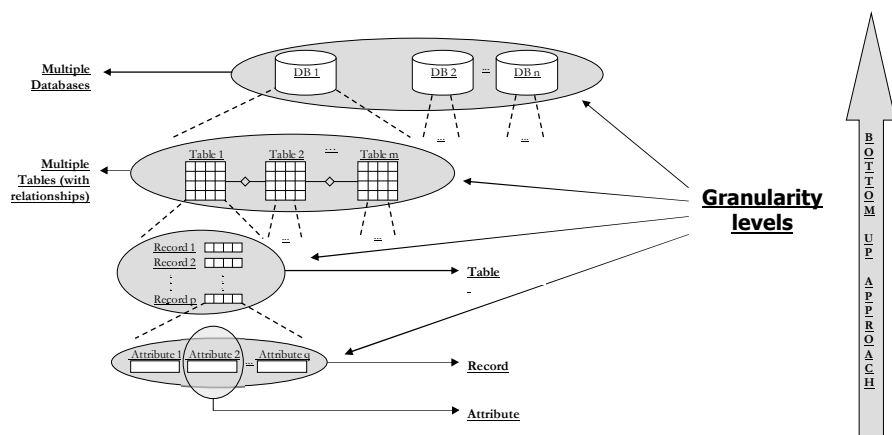
Concept of DQP

- Problems related with the data quality of the attributes' values
- Data values are affected by different kinds of quality problems (errors, anomalies, or *dirty*ness)
- DQPs arise in:
 - single data source
 - data migration
 - integration of multiple data sources
 - data-based projects
 - data warehouses
 - data mining
- Important due to the *Garbage In Garbage Out* (GIGO) principle

3

3

Approach to Identify DQPs












- Based on this data organization model
- Identification of all the DQPs that can be found in each granularity level

4

4

DQPs in a Single Attribute of a Single Record

I	a_1	a_2	=	a_m
r_1			=	
r_2			=	
\vdots	\vdots	\vdots	\vdots	\vdots
r_n			=	

5

5

Missing Value

	at_1	at_2	name	...	at_m
r_1	xxx	xxx	Carl Louis	...	xxx
r_2	xxx	xxx		...	xxx
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
r_n	xxx	xxx	Ben Johnson	...	xxx

6

6

Syntax Violation

	at ₁	at ₂	order_date	...	at _m
r ₁	xxx	xxx	24/10/2012	...	xxx
r ₂	xxx	xxx	2012/10/26	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	27/10/2012	...	xxx

7

7

Domain Violation

	at ₁	at ₂	ordered_quantity	...	at _m
r ₁	xxx	xxx	4	...	xxx
r ₂	xxx	xxx	-1	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	1	...	xxx

8

8

Misspelling Error

	at ₁	at ₂	city	...	at _m
r ₁	xxx	xxx	New York	...	xxx
r ₂	xxx	xxx	Bostom	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	Washington	...	xxx

9

9

Overloaded Attribute

	at ₁	at ₂	name	...	at _m
r ₁	xxx	xxx	Bill Clinton	...	xxx
r ₂	xxx	xxx	Dr. Barack Obama	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	George Bush	...	xxx

10

10

Incomplete Value

	at ₁	at ₂	address	...	at _m
r ₁	xxx	xxx	Sun Street, 123	...	xxx
r ₂	xxx	xxx	Flowers Street	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	Baker Street, 321	...	xxx

11

11

Wrong Value

	at ₁	at ₂	Marital Status	...	at _m
r ₁	xxx	xxx	Married	...	xxx
r ₂	xxx	xxx	Single	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	Divorced	...	xxx

12

12

DQPs in a Single Attribute in Multiple Records

T	a_1	a_2	\dots	a_m
r_1	■	□	\vdots	□
r_2	■	□	\vdots	□
\vdots	\vdots	\vdots	\vdots	\vdots
r_n	■	□	\vdots	□

13

13

Uniqueness Violation

	Name	at_2	taxpayer_nr	\dots	at_m
r_1	George Clooney	xxx	196 567 931	...	xxx
r_2	Harrison Ford	xxx	187 323 436	...	xxx
r_3	Brad Pitt	xxx	205 239 894	...	xxx
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
r_n	Judy Foster	xxx	187 323 436	...	xxx

14

14

Existence of Synonyms

	at ₁	at ₂	job	...	at _m
r ₁	xxx	xxx	Researcher	...	xxx
r ₂	xxx	xxx	Professor	...	xxx
r ₃	xxx	xxx	Electrician	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	Teacher	...	xxx

15

15

Violation of Business Rule

	invoice_nr	at ₂	invoice_date	...	at _m
r ₁	20121100	xxx	25/10/2012	...	xxx
r ₂	20121101	xxx	24/10/2012	...	xxx
r ₃	20121102	xxx	25/10/2012	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮
r _n	20121178	xxx	05/11/2012	...	xxx

The values of *invoice_date* must appear by ascending order !

16

16

DQPs in Multiple Attributes of a Single Record

T	a ₁	a ₂	=	a _m
r ₁	■	■	=	■
r ₂	□	□	=	□
⋮	⋮	⋮	⋮	⋮
r _n	□	□	=	□

17

17

Violation of Business Rule

	at ₁	quantity	unit_price	total_prod	...	at _m	
r ₁	xxx	2	3	6	...	xxx	
r ₂	xxx	2	5	5	...	xxx	
r ₃	xxx	3	4	12	...	xxx	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
r _n	xxx	2	1	2	...	xxx	

total_prod must be equal to *quantity * unit_price*!

18

18

DQPs at the Single Table Level

T	a_1	a_2	\dots	a_m
r_1	■	■	\dots	■
r_2	■	■	\dots	■
\vdots	\vdots	\vdots	\vdots	\vdots
r_n	■	■	\dots	■

19

19

Violation of Functional Dependency

	at_1	zip_code	city	\dots	at_m
r_1	xxx	4000	Oporto	\dots	xxx
r_2	xxx	4000	Lisbon	\dots	xxx
r_3	xxx	1000	Lisbon	\dots	xxx
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
r_n	xxx	4000	Oporto	\dots	xxx

20

20

Duplicate Records (Equal)

	id	name	address	taxpayer_nr	...	at _m
r ₁	xxx	xxx	xxx	xxx	...	xxx
r ₂	10	Cliff Barnes	Flowers Street, 123	205 239 894	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	72	Cliff Barnes	Flowers Street, 123	205 239 894	...	xxx

21

21

Duplicate Records (Approximate)

	id	name	address	taxpayer_nr	...	at _m
r ₁	xxx	xxx	xxx	xxx	...	xxx
r ₂	10	Cliff Barnes	Flowers Street, 123	205 239 894	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	72	C. Barnes	Flowers St., 123	205 239 894	...	xxx

22

22

Duplicate Records (Inconsistent)

	id	name	address	taxpayer_nr	...	at _m
r ₁	xxx	xxx	xxx	xxx	...	xxx
r ₂	10	Cliff Barnes	Flowers Street, 123	205 239 894	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	72	Cliff Barnes	Sun Street, 321	205 239 894	...	xxx

23

23

Violation of Business Rule

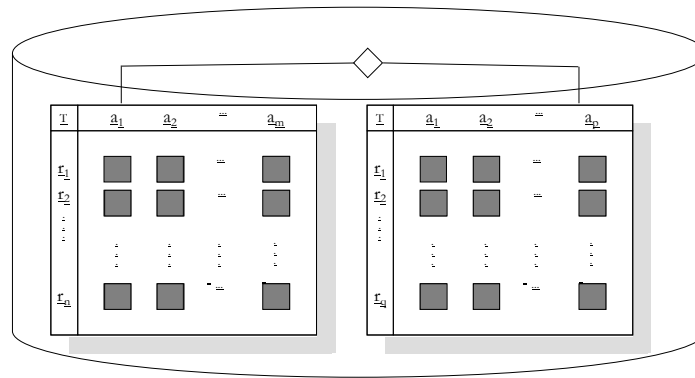
	at ₁	at ₂	...	at _m
r ₁	xxx	xxx	...	xxx
r ₂	xxx	xxx	...	xxx
⋮	⋮	⋮	⋮	⋮
r ₁₂	xxx	xxx	...	xxx

The number of product families (records) must not be superior to 10!

24

24

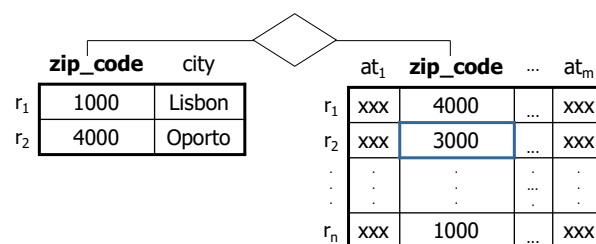
DQPs at the Level of Multiple Tables



25

25

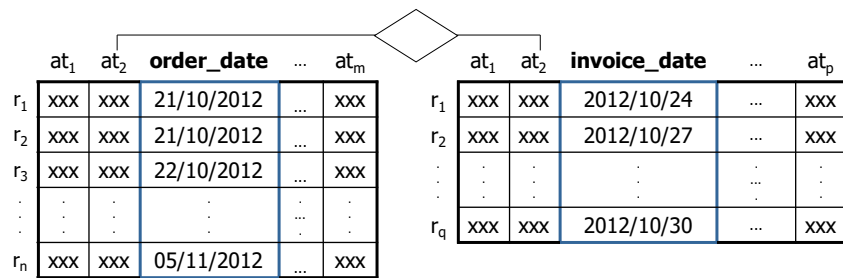
Referential Integrity Violation



26

26

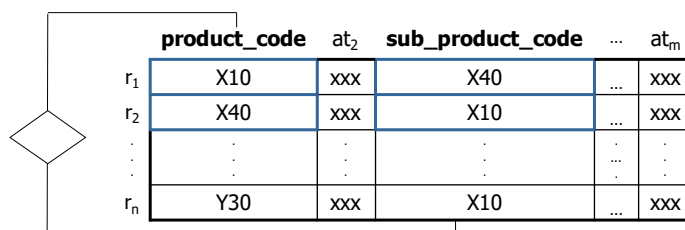
Heterogeneity of Syntaxes



27

27

Circularity among Tuples in a Self-Relationship

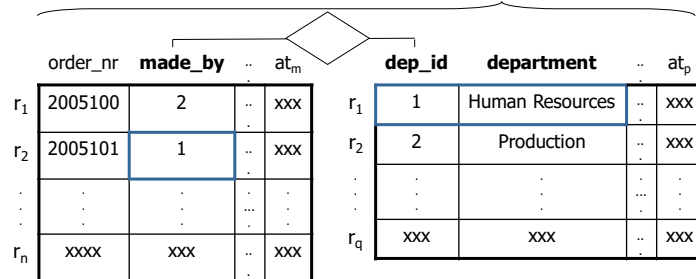


28

28

Violation of Business Rule

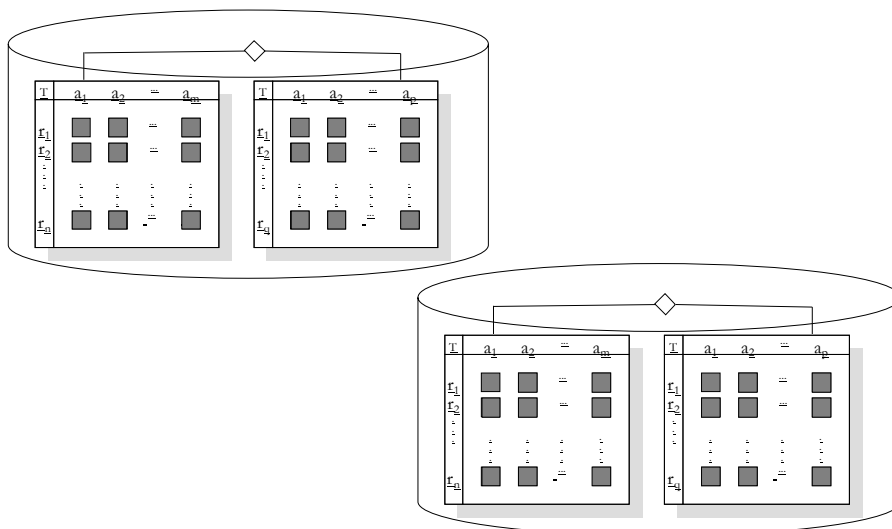
Orders can only be made by the provisions department or production department !



29

29

DQPs at the Level of Multiple Databases



30

30

Heterogeneity of Syntaxes

DB₁					DB₂				
	at ₁	order_date	...	at _m		at ₁	order_date	...	at _p
r ₁	xxx	21/10/2012	...	xxx	r ₁	xxx	2012/10/21	...	xxx
r ₂	xxx	21/10/2012	...	xxx	r ₂	xxx	2012/10/21	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	05/11/2012	...	xxx	r _q	xxx	2012/11/05	...	xxx

31

31

Heterogeneity of Measure Units

DB₁					DB₂				
	prod_id	unit_price	...	at _m		prod_id	unit_price	...	at _p
r ₁	xpto	5	...	xxx	r ₁	xpto	6.05	...	xxx
r ₂	ypto	12	...	xxx	r ₂	ypto	13.20	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	zpto	4	...	xxx	r _q	zpto	4.40	...	xxx

→ Dollars
← Euros

32

32

Heterogeneity of Domains

DB₁					DB₂				
	at ₁	gender	...	at _m		at ₁	gender	...	at _p
r ₁	xxx	M	...	xxx	r ₁	xxx	1	...	xxx
r ₂	xxx	F	...	xxx	r ₂	xxx	2	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	M	...	xxx	r _q	xxx	1	...	xxx

33

33

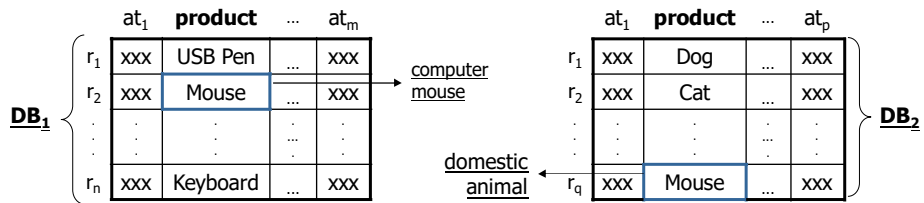
Existence of Synonyms

DB₁					DB₂				
	at ₁	job	...	at _m		at ₁	job	...	at _p
r ₁	xxx	Policeman	...	xxx	r ₁	xxx	Electrician	...	xxx
r ₂	xxx	Teacher	...	xxx	r ₂	xxx	Plumber	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	Researcher	...	xxx	r _q	xxx	Professor	...	xxx

34

34

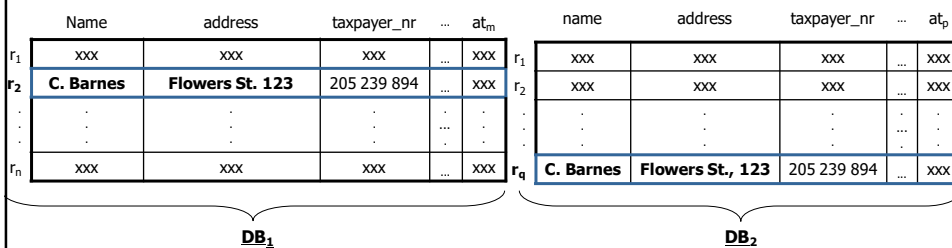
Existence of Homonyms



35

35

Duplicate Records (Equal)



36

36

Duplicate Records (Approximate)

	name	address	taxpayer_nr	...	at _m		name	address	taxpayer_nr	...	at _p
r ₁	xxx	xxx	xxx	...	xxx	r ₁	xxx	xxx	xxx	...	xxx
r ₂	Cliff Barnes	Flowers Street, 123	205 239 894	...	xxx	r ₂	xxx	xxx	xxx	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	xxx	...	xxx	r _q	C. Barnes	Flowers St., 123	205 239 894	...	xxx

DB₁ **DB₂**

37

37

Duplicate Records (Inconsistent)

	name	address	taxpayer_nr	...	at _m		name	address	taxpayer_nr	...	at _p
r ₁	xxx	xxx	xxx	...	xxx	r ₁	xxx	xxx	xxx	...	xxx
r ₂	Cliff Barnes	Flowers Street, 123	205 239 894	...	xxx	r ₂	xxx	xxx	xxx	...	xxx
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r _n	xxx	xxx	xxx	...	xxx	r _q	Cliff Barnes	Sun Street, 123	205 239 894	...	xxx

DB₁ **DB₂**

38

38

Violation of Business Rule

The maximum number of projects for a manager is two !

DB₁					DB₂				
	id_proj	manager_name	..	at _m		id_proj	manager_name	..	at _p
r ₁	XY100	A. Schwarzenegger	..	xxx	r ₁	AB900	Steven Segal	..	xxx
r ₂	YW200	Sylvester Stallone	..	xxx	r ₂	BC800	Sylvester Stallone	..	xxx
..
..
r _n	WZ300	Sylvester Stallone	..	xxx	r _q	CB700	Steven Segal	..	xxx

39

39

Reference

Paulo Oliveira, Fátima Rodrigues and Pedro Henriques – “A Formal Definition of Data Quality Problems”. In *Proceedings of the 10th International Conference on Information Quality*, MIT, Boston, EUA, November of 2005. p. 13-26.

40