Information Quality Techniques: a (Quick) Overview

Monica Scannapieco

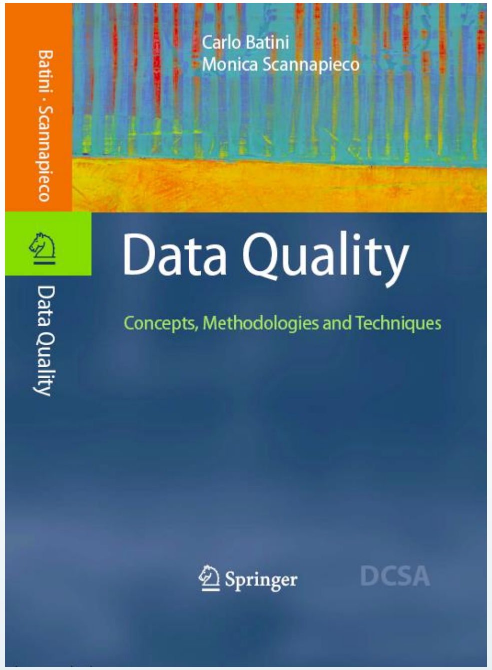
Italian National Institute of Statistics - Istat

Outline

- Information Quality: the concept and the (main) dimensions
- How to measure Information Quality

Information Quality vs. Data Quality

- Data quality is often referred to “structured data”, like e.g. a relational table
- However, a vast amount of realities is instead represented by types of information that are not structured data:
  - a photo of a landscape,
  - a map and a descriptive text in a travel guide,
  - newspaper articles,
  - satellite imagery etc.
- Information Quality is a more general concept than Data Quality and takes all types of information into account
- In the following we will than refer to Information Quality (IQ)



Reference for structured data

2006

Reference for data quality beyond structured data

2016

Information Quality: Beyond Accuracy

- When people think about IQ, they often reduce it just to accuracy, e.g., the city name "Chicago" misspelled as "Chcago"
- However, IQ is more than simply accuracy: other significant **dimensions** such as completeness, consistency, and currency are necessary in order to fully characterize it.

A relation Movies with IQ problems

Accuracy (swapped directors)

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	null	1964	0	1985

Accuracy

Consistency

Consistency

Completeness

Currency

Syntactic Accuracy= Distance functions for structured data

Type	Example
Identity	'Smith'= 'Smith'
Simple distance	'Smith' similar to 'Smth'
Complex distance (lot of distance functions: Hamming, bigrams, trigrams, etc.)	'Smith' similar to 'Smtih'
Acquisition process driven	Pain, Pane, Payn, Payne,etc. have a SoundexCode P500
Transformation	JFK Airport Acronym of John Fitzgerald Kennedy Airport

Semantic Accuracy

- Measures the distance of a represented value from the real world value
  - In the example, the swapped movie directors is a problem of semantic accuracy

Completeness

Completeness is defined as the degree to which a given data collection includes the data describing the corresponding set of real-world objects.
According to the type of structure considered, we may distinguish:

- Value completeness, to capture the presence of null values for some fields of a tuple;
- Tuple completeness, to characterize the completeness of a tuple with respect to the values of all its fields;
- Attribute completeness, to measure the number of null values of a specific attribute in a relation;
- Relation completeness, to capture the presence of null values in a whole relation.

@C.Batini, F. Cabitza, G. Pasi, R. Schettini, 2008

On completeness: The Person relation, with different null value meanings for the e-mail attribute

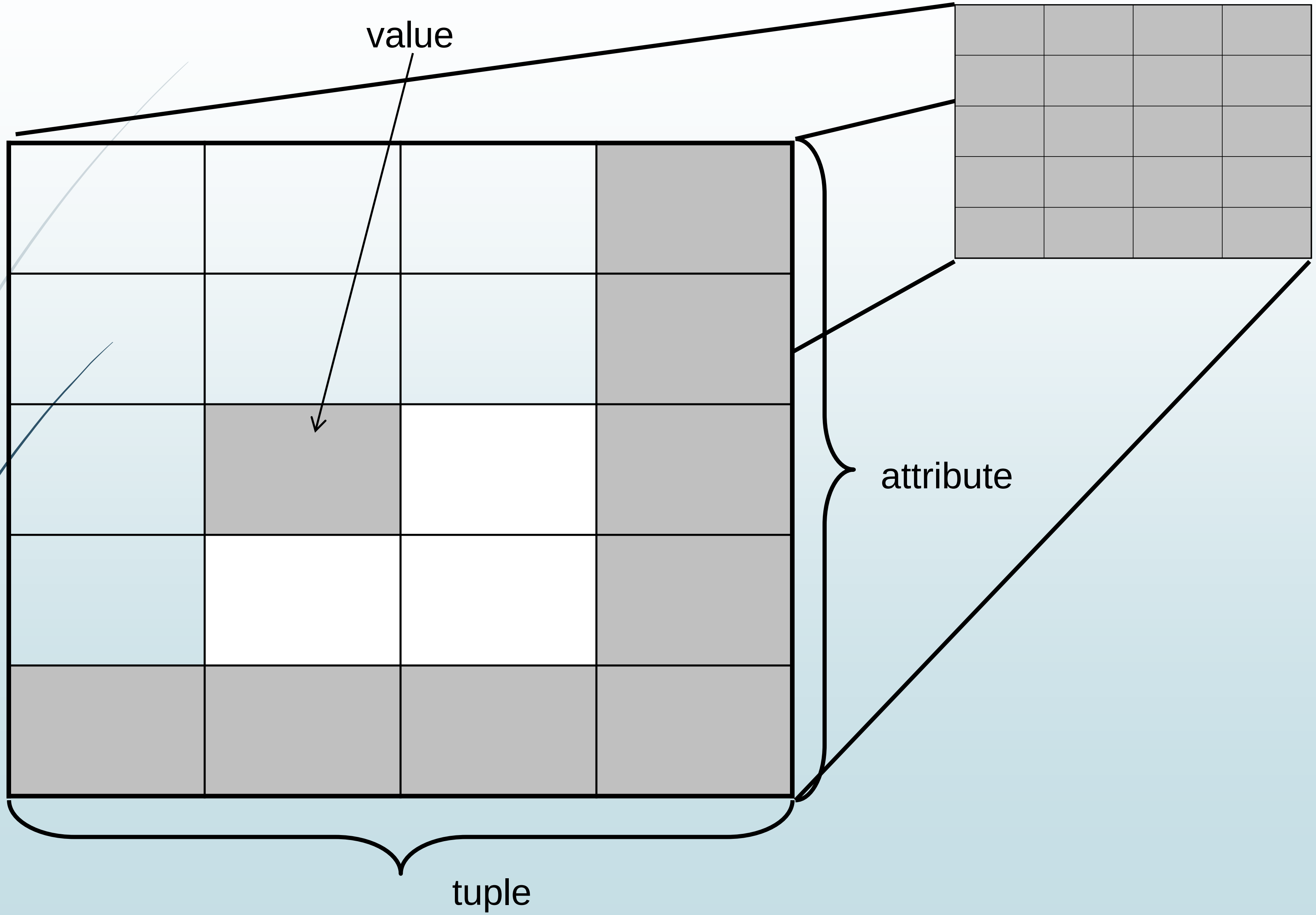
ID	Name	Surname	BirthDate	Email
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	NULL
3	Anthony	White	01/01/1936	NULL
4	Marianne	Collins	11/20/1955	NULL

not existing

existing but unknown

not known if existing

Completeness of different elements in the relational model



Consistency

- Captures the violation of semantic rules defined over a set of data items.
- Based on schema properties or integrity constraints or business rules
  - intrarelation constraints
    - regard single/multiple attributes of a relation.
  - interrelation constraints
    - involve attributes of more than one relation.
- Example: Last remake Year

Time-related dimensions: Timeliness, Currency, Volatility

Dimensions	Metrics Definition
Timeliness	Percentage of process executions able to be performed within the required time frame
Currency	Currency = Time in which data are stored in the system – time in which data are updated in the real world
Currency	Time of last update
Currency	Currency = Request time- last update
Currency	Currency = Age + (Delivery time- Input time)
Volatility	Length of time that data remain valid

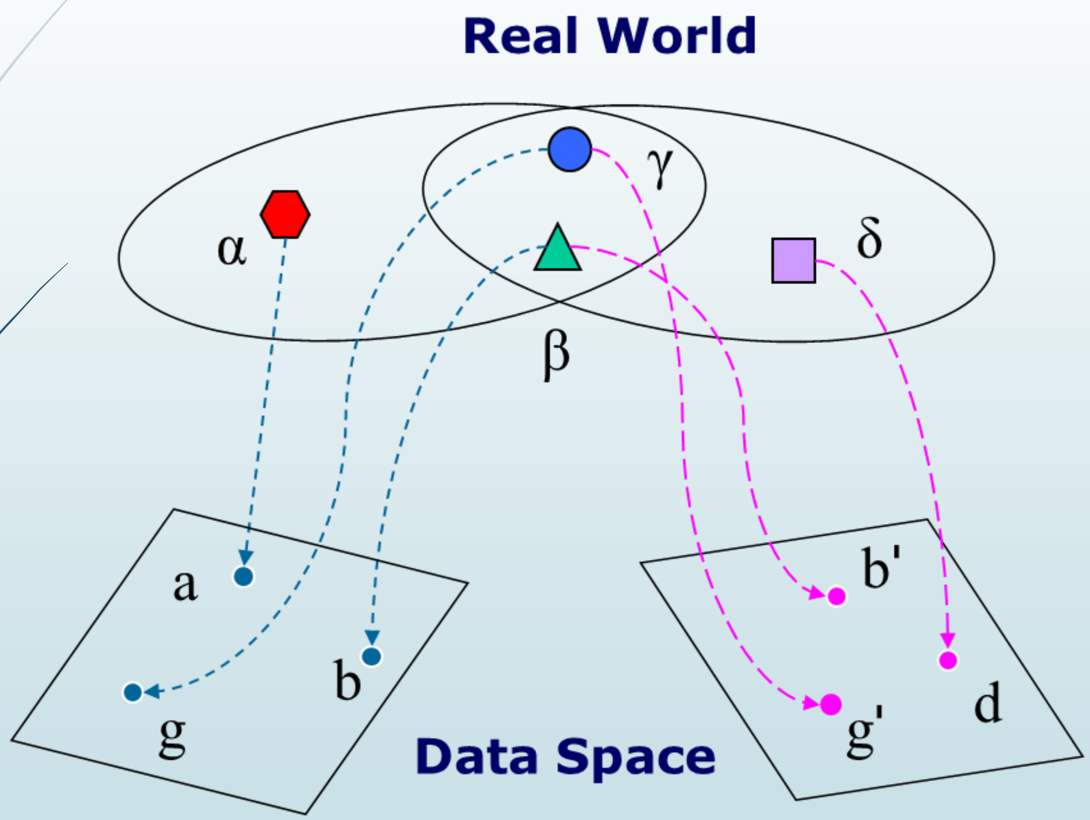
25012 ISO/IEC standard on data quality

DQ characteristic	Definition (all definitions except for completeness and accessibility begin with: the degree to which data has attributes that…")
Correctness	correctly represent the true value of the intended attribute of a concept or event in a specific context of use
Completeness	subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use
Consistency	are free from contradiction and are coherent with other data in a specific context of use
Credibility	are regarded as true and believable by users in specific context of use.
Currentness	are of the right age in a specific context of use
Accessibility	data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability
Compliance	adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use
Confidentiality	ensure that it is only accessible and interpretable by authorized users in a specific context of use
Efficiency	can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use
Precision	are exact or that provide discrimination in a specific context of use
Traceability	provide an audit trail of access to the data and of any changes made to the data in a specific context of use
Understandability	enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use
Availability	enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use
Portability	enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use
Recoverability	enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use

IQ Activities

- Standardizazion (Normalization)
- Object Identification (Record Linkage, Record Matching, Entity Resolution)
- Data integration with Schema Reconciliation & Instance-level Conflict Resolution
- Source Trustworthiness
- Quality Composition
- Error Localization
- Error Correction
- Cost optimization (quality-cost trade off)
- Applying methodologies for DQ assessment and improvement

Object Matching



OM Problem

a ↔ b'	no
a ↔ g'	no
a ↔ d	no
b ↔ b'	yes
…………	…
…………	…
g ↔ g'	yes
g ↔ d	no

How three agencies represent the same business

Agency	Identifier	Name	Type of activity	Address	City
Agency 1	CNCBTB765SDV	Meat production of John Ngombo	Retail of bovine and ovine meats	35 Niagara Street	New York
Agency 2	0111232223	John Ngombo canned meat production	Grocer's shop, beverages	9 Rome Street	Albany
Agency 3	CND8TB76SSDV	Meat production in New York state of John Ngombo	Butcher	4, Garibaldi Square	Long Island

Object Matching

- The three records present several differences:

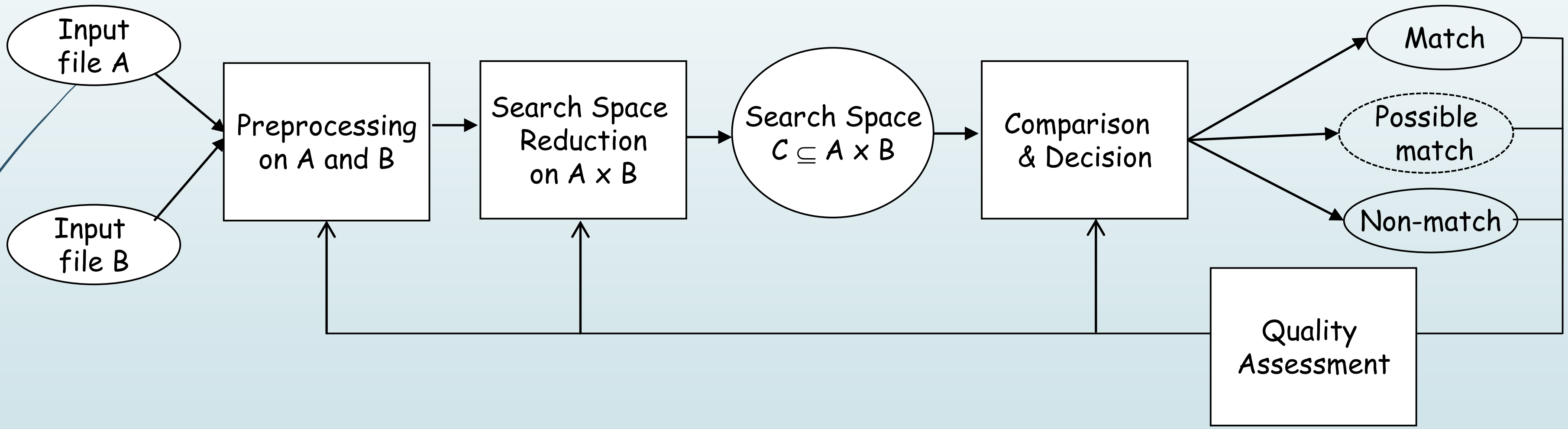
1. Different values for identifiers, due to different coding policies. Even if coding policies are the same, there can be errors (see 1 e 3)
2. Names are different, even if there are several common parts
3. Activity types are different due to several possible causes: errors, false declarations, updates on different dates…
4. Further differences on Address e City

- Object Identification is an IQ activity aiming to identify if different representations of an entity in an information system, are (or not) the same real world entity

Assessing and Improving IQ in data integration contexts: Object Matching

- OM is a crucial step to integrate data at istance level
- RISIS setting does requires data integration
- Possible methodological solution for addressing quality issues of RISIS

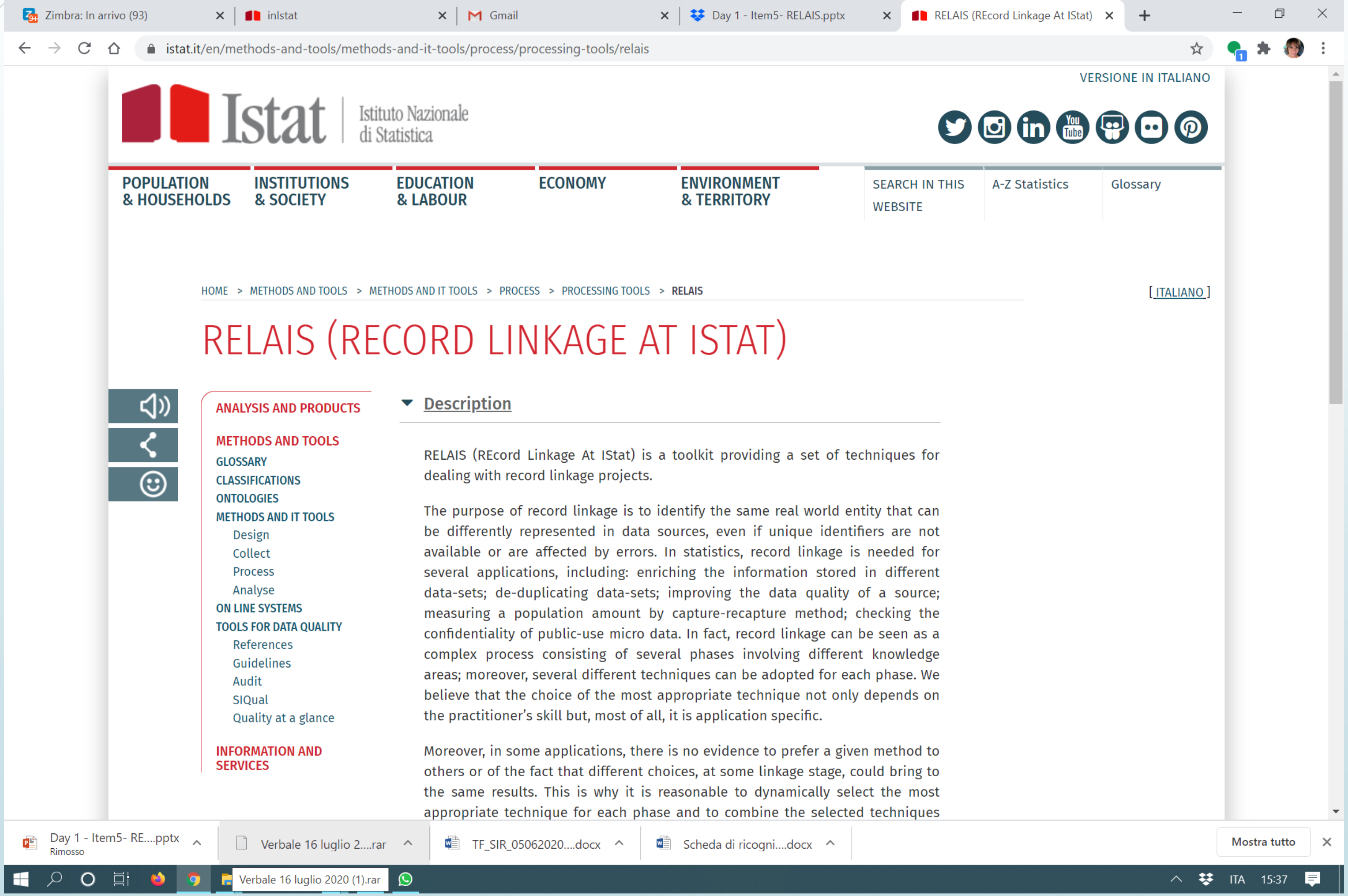
Relevant steps of object identification techniques



Object identification techniques (some examples)

Name	Technical Area	Type of data
Fellegi and Sunter and extensions	probabilistic	Two files
Cost-based	probabilistic	Two files
Sorted Neighborhood and variants	empirical	Two files
Delphi	empirical	Two relational hierarchies
DogmatiX	empirical	Two XML documents
Intelliclean	knowledge-based	Two files
Atlas	knowledge-based	Two files

An Example: Object Matching at work



VERSIONE IN ITALIANO

Istat | Istituto Nazionale di Statistica

POPULATION & HOUSEHOLDS | INSTITUTIONS & SOCIETY | EDUCATION & LABOUR | ECONOMY | ENVIRONMENT & TERRITORY | SEARCH IN THIS WEBSITE | A-Z Statistics | Glossary

HOME > METHODS AND TOOLS > METHODS AND IT TOOLS > PROCESS > PROCESSING TOOLS > RELAIS

[ITALIANO]

RELAIS (RECORD LINKAGE AT ISTAT)

ANALYSIS AND PRODUCTS

METHODS AND TOOLS
- GLOSSARY
- CLASSIFICATIONS
- ONTOLOGIES
- METHODS AND IT TOOLS
  - Design
  - Collect
  - Process
  - Analyse
- ON LINE SYSTEMS
- TOOLS FOR DATA QUALITY
  - References
  - Guidelines
  - Audit
  - SIQual
  - Quality at a glance

INFORMATION AND SERVICES

Description

RELAIS (REcord Linkage At IStat) is a toolkit providing a set of techniques for dealing with record linkage projects.

The purpose of record linkage is to identify the same real world entity that can be differently represented in data sources, even if unique identifiers are not available or are affected by errors. In statistics, record linkage is needed for several applications, including: enriching the information stored in different data-sets; de-duplicating data-sets; improving the data quality of a source; measuring a population amount by capture-recapture method; checking the confidentiality of public-use micro data. In fact, record linkage can be seen as a complex process consisting of several phases involving different knowledge areas; moreover, several different techniques can be adopted for each phase. We believe that the choice of the most appropriate technique not only depends on the practitioner's skill but, most of all, it is application specific.

Moreover, in some applications, there is no evidence to prefer a given method to others or of the fact that different choices, at some linkage stage, could bring to the same results. This is why it is reasonable to dynamically select the most appropriate technique for each phase and to combine the selected techniques

Principal Functionalities

- Input/Output Management
  - Back-up support
  - Residuals management
- Data Profiling
  - Metadata for blocking variable selection
  - Metadata for matching variable selection
- Search Space Creation and Reduction methods
  - Cross Product
  - Blocking
  - Sorted Neighborhood
  - Nested blocking
  - Simhash (new!)

Principal Functionalities

- Comparison Functions
- Deterministic Decision Models
  - Exact
  - Rule-based
- Probabilistic Decision Model
  - Fellegi-Sunter
- 1: 1 Reduction methods
  - Optimal
  - Greedy

Data Profiling

- Blocking and matching variables:
  - Completeness
  - Accuracy
  - Consistency
  - Categories
  - Frequency distribution
  - Entropy
- Blocking
  - Blocking adequacy
- Matching
  - Correlation

Search Space Creation and Reduction Methods

- Search Space Creation
  - Cross product
- Reduction Methods: Blocking
  - Selection of a blocking key (two or more variables)
  - Block modality table reports information on created blocks (sizes, number of blocks,…)
- Reduction Method: Sorted Neighborhood Method (SNM)
  - Selection of a sorting key (two or more variables)
  - Choice of the window size

Comparison Functions

- Equality
- Jaro
- Dice
- Jaro-Winkler
- Levenshtein
- 3-grams
- Soundex
- Numeric Comparison

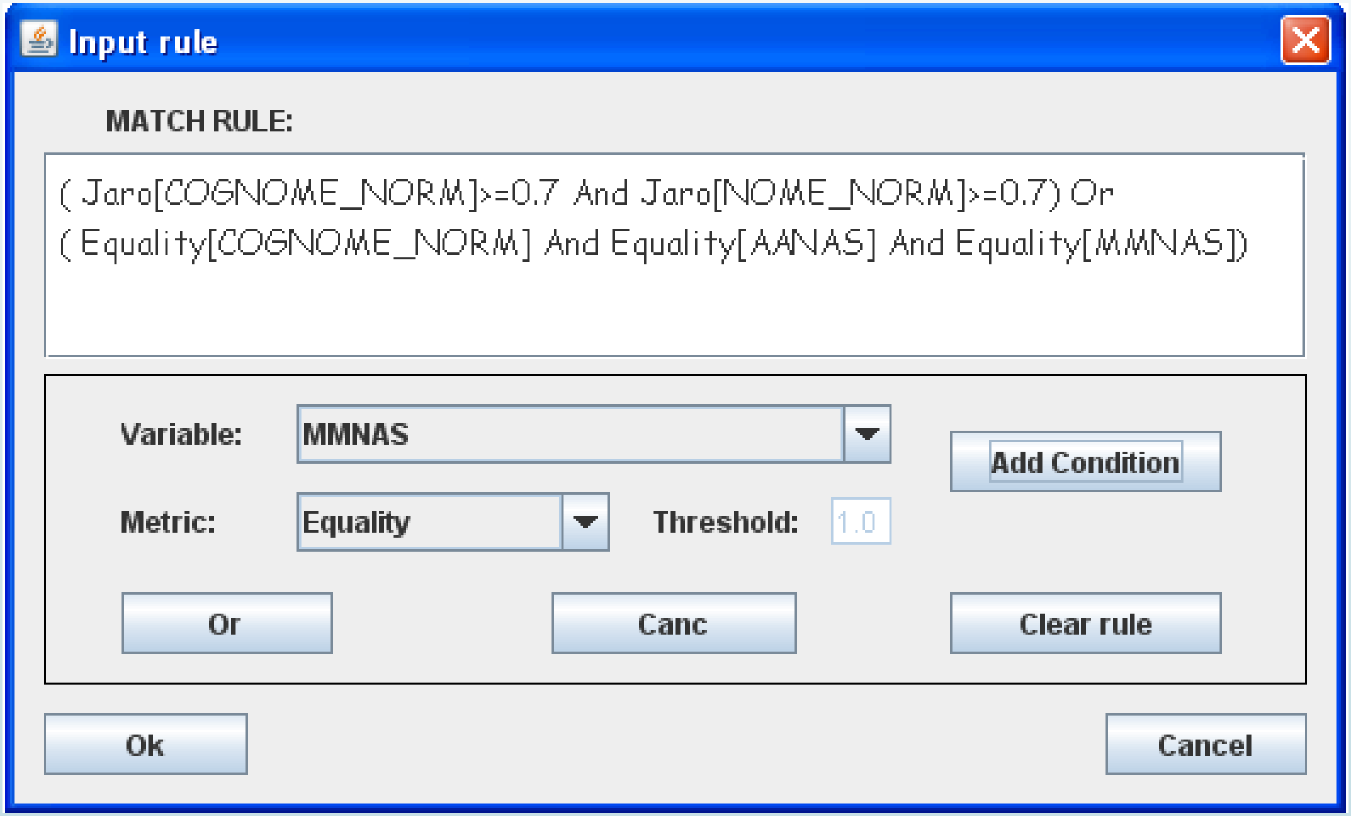
Deterministic Decision Models: Equality Match

- Exact matching: relational JOIN over specified variables
- Useful at the initial stage of a RL process, when it makes sense to "prune" the pairs to compare by removing exact matches

Deterministic Decision Models: Rule Based

- Rules specified through a GUI
- The GUI allows specification of
  - Variables of the formula
  - Comparison functions
  - Thresholds for each variable
  - Operators to combine atomic formulas (AND/OR)

Deterministic Decision Models: Rule Based



Probabilistic Decision Model: Fellegi-Sunter

- Steps:
  1. Choice of matching variables
  2. Choice of comparison functions and thresholds (for each variable)
  3. Contingency table computation
  4. EM Estimation → MU table result

estimates of frequency distributions

Posterior probability f_m/(f_m+f_u)

MU_TABLE

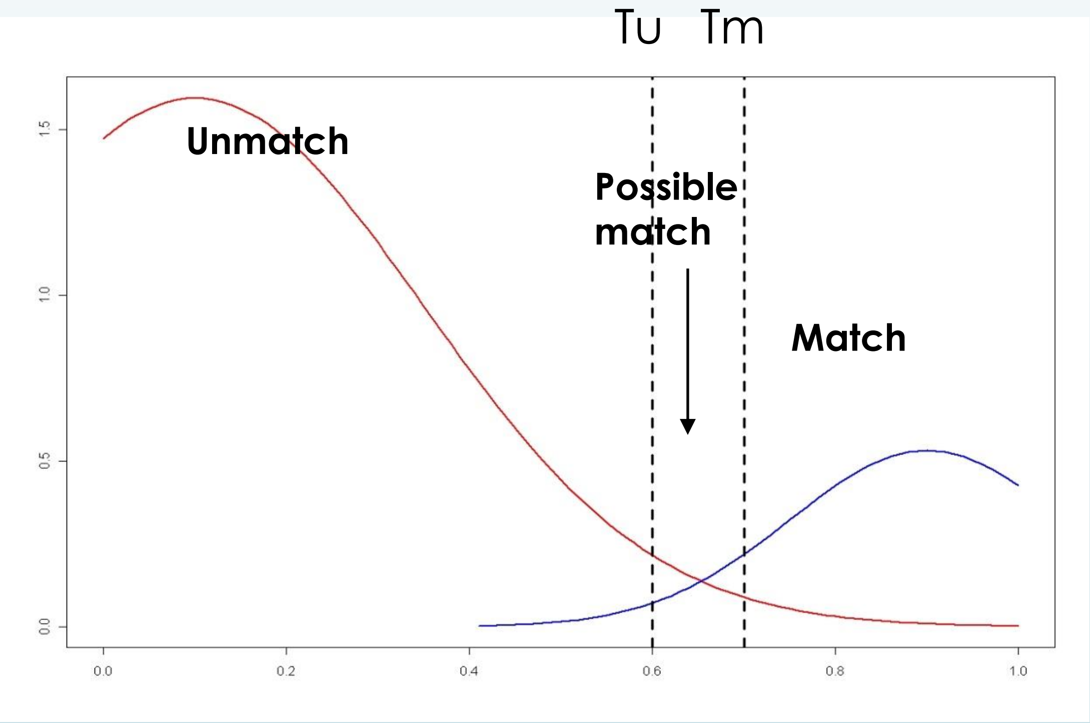
business_name	city	classification	year_begin	f_m	f_u	m	u	r	p_post
0	0	0	0	3482.55459	128177.44...	0.39638	0.84771	0.46759	0.02645
0	1	0	0	588.45339	8781.54661	0.06698	0.05808	1.15323	0.0628
0	0	0	1	313.26504	2301.73496	0.03566	0.01522	2.34224	0.1198
0	0	1	0	2405.43876	11017.56124	0.27378	0.07287	3.75737	0.1792
0	1	0	[illegible]	[illegible]	[illegible]	0.00475	8.2E-4	5.77679	0.25131
0	1	1	[illegible]	[illegible]	[illegible]	0.04099	0.00442	9.26699	0.35
0	0	1	[illegible]	[illegible]	[illegible]	0.01552	8.2E-4	18.8215	0.52237
0	1	1	1	24.07464	8.92536	0.00274	6.0E-5	46.42042	0.72953
1	0	0	0	675.99984	1.6E-4	0.07694	0.0	7.5008830...	1.0
1	1	0	0	58.99999	1.0E-5	0.00672	0.0	1.8499804...	1.0
1	0	0	1	36.0	0.0	0.0041	0.0	3.7573590...	1.0
1	0	1	0	446.99999	1.0E-5	0.05088	0.0	6.0274678...	1.0
1	1	0	1	30.0	0.0	0.00341	0.0	9.2669631...	1.0
1	1	1	0	67.0	0.0	0.00763	0.0	1.4865846...	1.0
1	0	1	1	30.0	0.0	0.00341	0.0	3.0192926...	1.0
1	1	1	1	89.0	0.0	0.01013	0.0	7.4466329...	1.0

Precision = TP/TP+FP and Recall=TP/TP+FN

Probabilistic Decision Model: Fellegi-Sunter

- Steps:
  1. Choice of matching variables
  2. Choice of comparison functions and thresholds (for each variable)
  3. Contingency table computation
  4. EM Estimation → MU table result
  5. Threshold Selection

Threshold Selection



1:1 Matching

DETERMINISTIC	PROBABILISTIC
LP Problem (Global Optimization)	LP Problem (Global Optimization)
Greedy Reduction	Greedy Reduction
<u>Subrules weight</u>	<u>R weight</u>

1:1 Matching Deterministic Model

- LP problem with input matrix:
  - Weight associated to each atomic subrule
  - Sum of weights
- Greedy: sorting of pairs by the sum of weights
  - Choices are local

(Current) RELAIS Download

- RELAIS released at:
  - ISTAT: https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/relais
  - Joinup: https://joinup.ec.europa.eu/collection/statistics/solution/relais-record-linkage-istat

Conclusions

- Information characterized by several dimensions
- A methdological approach to assess and improve IQ in RISIS: object matching
- Object matching at work: Relais