Seminar on Record Linkage

Nicoletta Cibella and Tiziana Tuoto

Italian National Statistical Institute - Istat

Integration, Quality, Research and Production Networks Development Department

Programme

Monday 5 November:

- 9:00-10:30 Definition of record linkage. Examples and motivations. Record linkage phases (Tiziana Tuoto)
- 10:30-10:50 Coffee break
- 10:50-13:00 Preprocessing and standardization. Blocking procedures. Type of matching. (Nicoletta Cibella)
- 13:00-14:00 Lunch break
- 14:00-15:30 RELAIS: a statistical toolkit for record linkage
- 15:30-15:50 Tea break
- 15:50-17:15 Applications of record linkage procedures and case studies

Programme

Thusday 6 November:

- 9:00-10:30 Types of record linkage methods. A statistical model for record linkage: the Fellegi-Sunter approach. (Nicoletta Cibella)
- 10:30-10:50 Coffee break
- 10:50-13:00 Outlines on possible extensions of the Fellegi-Sunter approach and estimation of the components of decision rules.Quality in record linkage. Outline on the analysis of linked data files. (Tiziana Tuoto)
- 13:00-14:00 Lunch break
- 14:00-16:00 Applications of record linkage procedures and case studies
- Course evaluation

Introduction to micro-data integration

Tiziana Tuoto

Italian National Statistical Institute - Istat

Integration, Quality, Research and Production Networks Development Department

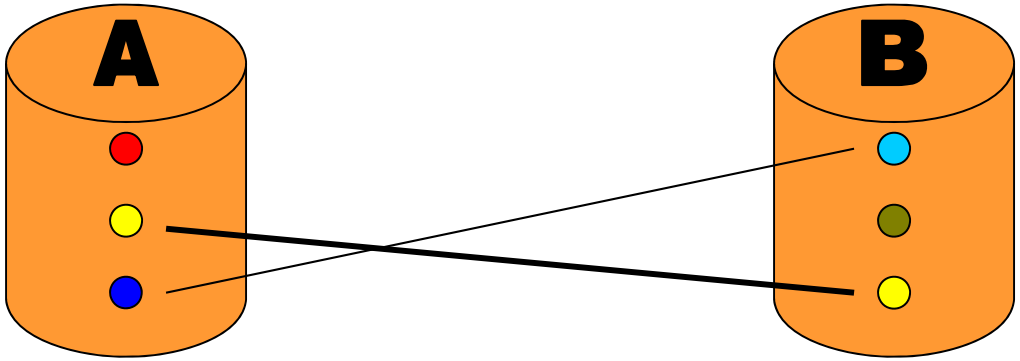
tuoto@istat.it

Micro-data integration

Record Linkage

Let A and B be two sets (sources) of data.

Aim: “integration” of the two sources : i.e. we want to identify records referred to the same entity, but belonging to different files, by means of “not perfetc” common key



Data sources

A: consists of n_A records

B: consists of n_B records

Some of the variables (X) are observed in both A and B (common variables)

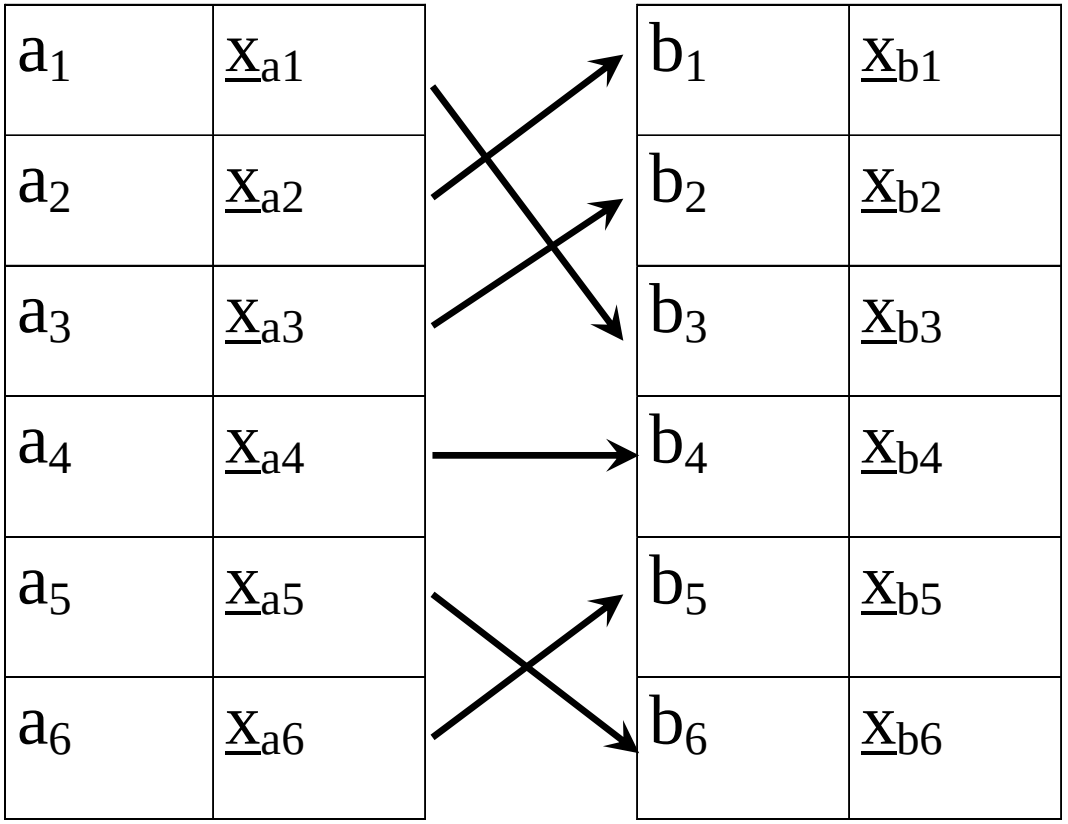
A	
X	Y
x_{a_1}	y_{a_1}
x_{a_2}	y_{a_2}
…	…
x_{an_A}	y_{an_A}

B	
X	Z
x_{b_1}	z_{b_1}
x_{b_2}	z_{b_2}
…	…
x_{bn_B}	z_{bn_B}

Objective: micro

Record detection is performed by means of the common variables X

There exist different kinds of integration at the micro level



Statistical methods for integration

It is useful to differentiate the methods according to the available input (i.e. the data sets to integrate)

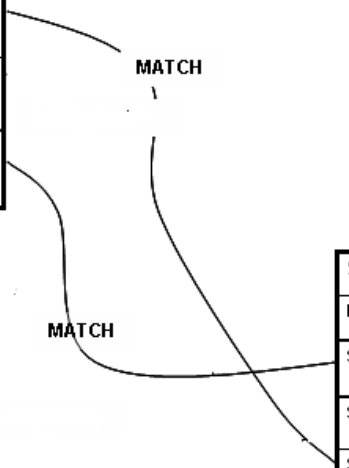
Input	**Method**
Two data sets that observe (partially) overlapping groups of units	Record linkage
Two independent samples, without any units in common	Statistical matching

Record linkage

Input: two data sets that observe overlapping groups of units

S1: SHOE SHOPS (OHIO)

ID	Name	Address	Telephone
S1.1	Rugged Boot The	4901 W Broad St Prairie Twp OH 43228	614-878-0569
S1.2	Rugged Boot The	4788 Columbus Park Lewis Center OH 43035	740-548-7463
S1.3	Springshod Footware	2300 E Kemper Rd Sharonville OH 45241	513-771-1175



S2: SHOE SHOPS (OHIO)

ID	Name	Address	Telephone
S2.1	Springshod Footware	2300 E Kemper Rd Cincinnati OH 45241-6501	513-771-1175
S2.2	Springshod Footware	8969 Kingsridge Drive Dayton, OH	937-312-0506
S2.3	The Rugged Boot	4901 W Broad St Columbus, OH	800-605-2668

Problem:

Lack of a unique and correct identifier

Solutions: presence of a set of variables that (jointly) allows the detection of records

Attention: variables can have “problems”!

Aim: highest number of correct linkages, lowest number of wrong linkages

Examples and motivations

Tiziana Tuoto

Italian National Statistical Institute - Istat

Integration, Quality, Research and Production Networks Development Department

tuoto@istat.it

Why record linkage?

According to Fellegi (1997)*, the development of tools for data integration is due to the intersection of these facts:

- occasion: construction of big data bases
- tool: computer
- need: new informative needs

*Fellegi (1997) “Record Linkage and Public Policy: A Dynamic Evolution”. In Alvey, Jamerson (eds) *Record Linkage Techniques*, Proceedings of an international workshop and exposition, Arlington (USA) 20-21 March 1997.

Objectives

Some objectives for record linkage

1. To have joint information on two or more variables observed in distinct data sources
2. To “enumerate” a population
3. To substitute (parts of) surveys with archives
4. To create a “list” of a population
5. Other official statistical objectives (to study the risk of identification of the released micro data)

Example 1 – joint info from different data sources

Problem: to analyze jointly the “risk factors” with the event “death”.

A)The risk factors are observed on ad hoc surveys (e.g. those on nutrition habits, work conditions, etc.)

B)The event “death” (after some months the survey is conducted) can be taken from administrative archives

These two sources (survey on the risk factors and death archive) should be “fused” so that each unit observed in the risk factor survey can be associated with a new dichotomous variable (equal to 1 if the person is dead and zero otherwise).

Example 1 – joint info from different data sources

Problem: integrate surveillance of road traffic accidents.

A) information on the circumstances of traffic accident and characteristic of roads and vehicles are reported by local policy authority on ad hoc survey

B)medical and socio-demographic data for killed people in road accidents can be taken from administrative archives (e.g. Italian National Vital Statistics Death Registry on causes of death)

These two sources have been “fused” so to enrich and update the systems of indicators on public health and security.

Example 1 – other analyses

Other important analyses on linked data sets performed in recent times:

Linked longitudinal employer-employee data set (Statistics New Zealand) based on linking administrative data held in the NZ Inland Revenue Department's tax system and Statistics new Zealand's list of NZ businesses (analysis of job and worker flows, employment tenure, multiple jobholding and business demography)

Example 1 – other analyses

Statistical longitudinal census dataset (ABS) An important feature of the Census Data Enhancement (CDE) project is the formation of a Statistical Longitudinal Census Dataset (SLCD) by bringing together data from the 2006 Census with data from the 2011 Census and future Censuses to build a picture of how society moves through various changes: which groups are affected by different types of change and in what way. The non-identifying grouped numeric code will be used in conjunction with characteristics such as age, sex, geographic region and country of birth to link records from the 5% SLCD to the 2016 Census and future Censuses using probabilistic record linkage techniques. Name and address information will not be used in the linkage process and will not be available for the 5% SLCD dataset as they are deleted at the end of Census processing.

Example 2 – to enumerate a population

Problem: what is the number of residents in Italy?

Often the number of residents is found in two steps, by means of a procedure known as “capture-recapture”. This method is usually applied to determine the size of animal populations.

A) Population census

B) Post enumeration survey (some months after the census) to evaluate Census quality and give an accurate estimate of the population size

USA - in 1990 *Post Enumeration Survey*, in 2000 *Accuracy and Coverage Evaluation*

Italy - in 2001 and 2011 (work in progress)

Example 2 – to enumerate a population

The result of the comparison between Census and post enumeration survey is a 2×2 table:

	Observed post enumer. survey	Non observed post enumer. survey
Observed in Census	n_{oo}	n_{on}
Non observed in Census	n_{no}	??

Example 2 - to enumerate a population

For short, for any distinct unit it is necessary to understand if it was observed

1) both in the census and in the PES
2) only in the census
3) only in the PES

These three values allow to estimate (with an appropriate model) the fourth value.

Example 3 – surveys and archives

Problem: is it possible to use jointly administrative archives and sample surveys?

At the micro level this means: to modify the questionnaire of a survey dropping those questions that are already available on some administrative archives (reduction of the response burden)

E.g., for enterprises:

Social security archives, chambers of commerce, …

Example 3 – surveys and archives

Example: built up an integrated system on the pregnancy outcome in Italy

At the micro level this means to link:

- administrative data on births,
- medical data on hospital admission forms for newborns, stillborns, deliveries, voluntary abortions, miscarriages
- surveys on newborns and mothers, voluntary abortions, miscarriages

The system presents a complete picture on pregnancy outcomes and allows to measure several relationships and indicators. Moreover it provides the join comparison of a great amount of data sources, useful to assess the level of quality and reliability of each one.

Example 4 – Creation of a list

Problem: what is the set of the active enterprises in Italy?

In Istat, ASIA (Archivio Statistico delle Imprese Attive) is the most important example of a creation of a list of units (the active enterprises in a time instant) “fusing” different archives.

It is necessary to pay attention to:

- Enterprises which are present in more than one archives (deduplication)
- Non active enterprises
- New born enterprises
- Transformations (that can lead to a new enterprise or to a continuation of the previous one)

Example 5 - Privacy

Problem: does it exist a “measure” of the degree of identification of the released microdata?

In order to evaluate if a method for the protection of data disclosure is good, it is possible to compare two datasets (the true and the protected ones) and detect how many modified records are “easily” linked to the true ones.

Other experiences?

In your opinion, is there any other record linkage objective that can be included in the previous list?

Real life examples are welcome!

Common aspects

- The previous examples compare two (or more) data sets (sources, archives, surveys,...)
- The objective is always micro: to link records in the two files that are available in the two data sets

Specific aspects

- Creation of a list: the procedure stops at the unit identification step
- Analysis, enumeration of a population: unit identification is only a step for the estimation of “aggregates”
- Imputation, privacy protection, analysis, enumeration of a population: if the variables used for linking the records (common variables X) are “unstable”, a fundamental role is played by the “probability of correct link”

ESSnets on data integration

ESSnet "Integration of surveys and administrative data“ 2006-2008

http://www.essnet-portal.eu/project-information/isad-finished

ESSnet “Data Integration“ 2009-2011

http://www.essnet-portal.eu/di/data-integration

On these sites it is possible to find

- Reports
- Experiences in the different ESS countries
- Links to web pages, software, publications for record linkage

(for statistical matching and micro integration processing, as well)

Definition of Record Linkage

Tiziana Tuoto

Italian National Statistical Institute - Istat

Integration, Quality, Research and Production Networks Development Department

tuoto@istat.it

Record Linkage

- T. Belin, D. Rubin (1995, JASA): The term “record linkage” refers to the use of algorithmic techniques for the record identification in different data bases that refer to the same unit.

- Synonyms: Exact Matching / Computer matching

(Gu et al.) While epidemiologists and statisticians speak of record linkage, the same process is called entity heterogeneity, entity identification, object isomerism, instance identification, merge/purge, entity reconciliation, list washing and data cleaning by computer scientists and others

Record Linkage – simple case

I have two data sets A and B (e.g. on individuals).

Every record possesses a unique unit identifier (e.g. a PIN) which is not affected by errors.

It is possible to neglect the other variables, and link the records with the same PIN.

Record Linkage – simple case

A					
Name	**Surname**	**ZIP code**	**Date of birth**	**Place of birth**	**PIN**
Mario	Rossi	00125	18/05/1970	Roma	RSSMRA70E18 H501T

B					
Name	**Surname**	**ZIP code**	**Date of birth**	**Place of birth**	**PIN**
					RSSMRA70E18 H501T

Record Linkage - intermediate case

A unique identifier does not exist, or cannot be used.

The other variables are jointly able to identify the unit (together they play the role of the identifier). Furthermore they are not affected by errors or missing items.

Record Linkage - intermediate case

A					
Name	**Surname**	**ZIP code**	**Date of birth**	**Place of birth**	**PIN**
Mario	Rossi	00125	18/05/1970	Roma	RSSMRA70E18 H501T

B					
Name	**Surname**	**ZIP code**	**Date of birth**	**Place of birth**	**PIN**
Mario	Rossi	00125	18/05/1970	Roma	---

Record Linkage – difficult case

A unique identifier does not exist, or cannot be used.

The other variables are jointly able to identify the unit (together they play the role of an identifier). Anyway there can be differences in the answers for:

- Errors
- Missing answers
- Correct answers with a different codification/structure
- Changes due to time.

Record Linkage – difficult case

A					
Name	**Surname**	**ZIP code**	**Date of birth**	**Place of birth**	**PIN**
Mario	Rossi	00125	18/05/70	Roma	RSSMRA70E18 H501T

B					
Name	**Surname**	**ZIP code**	**Date of birth**	**Place of birth**	**PIN**
M.	Russi	00152	18 maggio 1970	Ostia	

Record Linkage - characteristics

- If a unique identifier exists, or if a set of variables not effected by errors can play the role of an identifier, the integration problem is straightforward.
- We consider the record linkage problem in the case there is not a unique identifier and the other variables can play the role of an identifier but are reported with error.

Record Linkage - characteristics

When a unique unit identifier is missing in at least one of the data sources to integrate, it is necessary to refer to the variables observed in the two files.

The problem is that these variables can be “unstable”:

1. Modification due to time changes (age, education, address,…)
2. Lack of accuracy in reporting or registering data
3. Different structure of some fields (address, date of birth)
4. Missing items

Record Linkage - characteristics

As a matter of fact, record linkage methods are due to:

1) Bad data “maintenance” (lack of a correct identifier – positive examples: Finland, Netherlands)

2) “pathology” in data (errors and modifications)

In this case, the problem is that

- Some true matches are not detected
- Some false matches are interpreted as true matches

Objective of a record linkage procedure: to minimize these linkage errors

Other experiences

- In your record linkage examples, do you have an identifier?
- Is the identifier “good” in both the data sets?
- If a unique identifier does not exist, what is the quality of the common variables in the data sets to link?

Record linkage Phases

Tiziana Tuoto

Italian National Statistical Institute - Istat

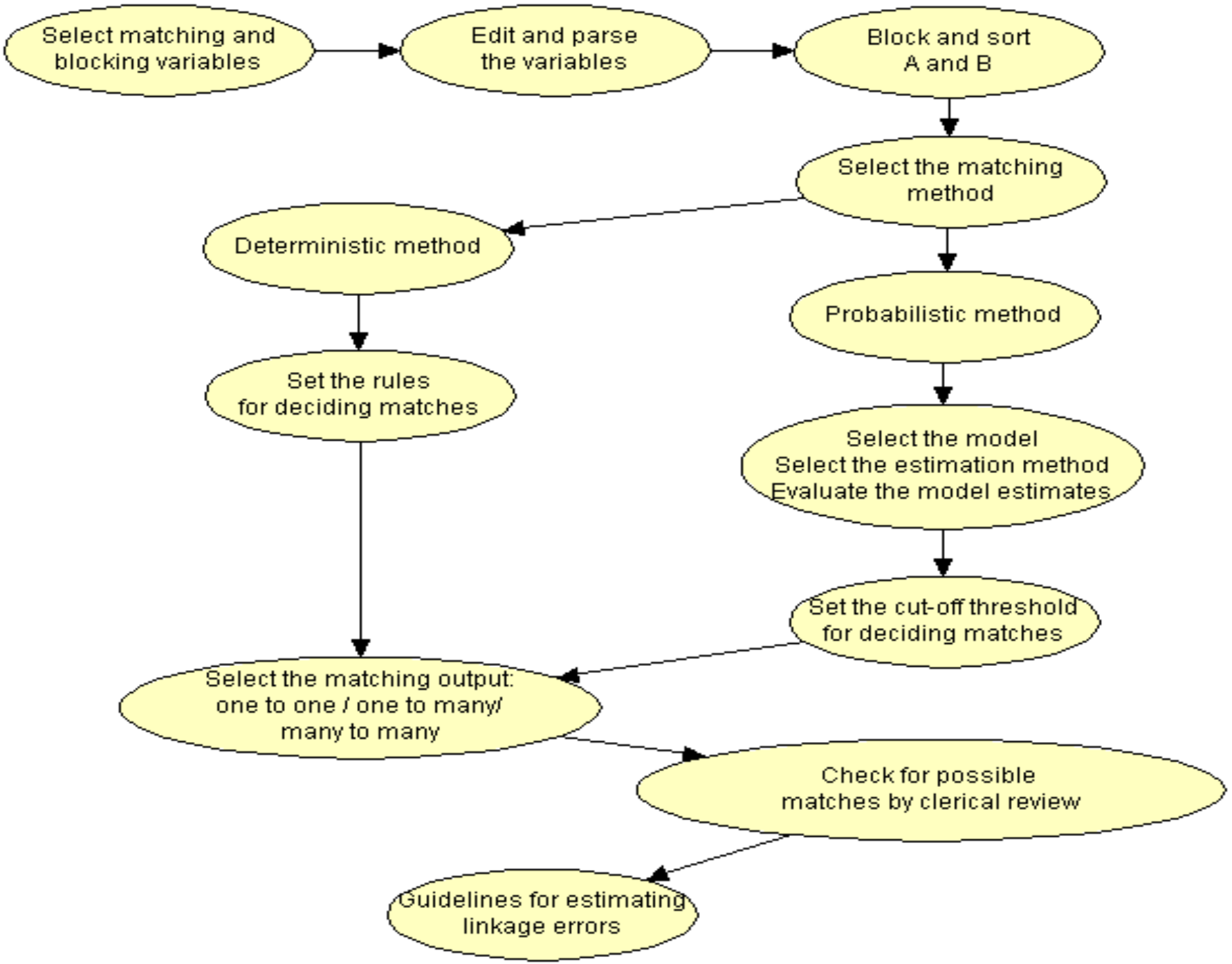
Integration, Quality, Research and Production Networks Development Department

tuoto@istat.it

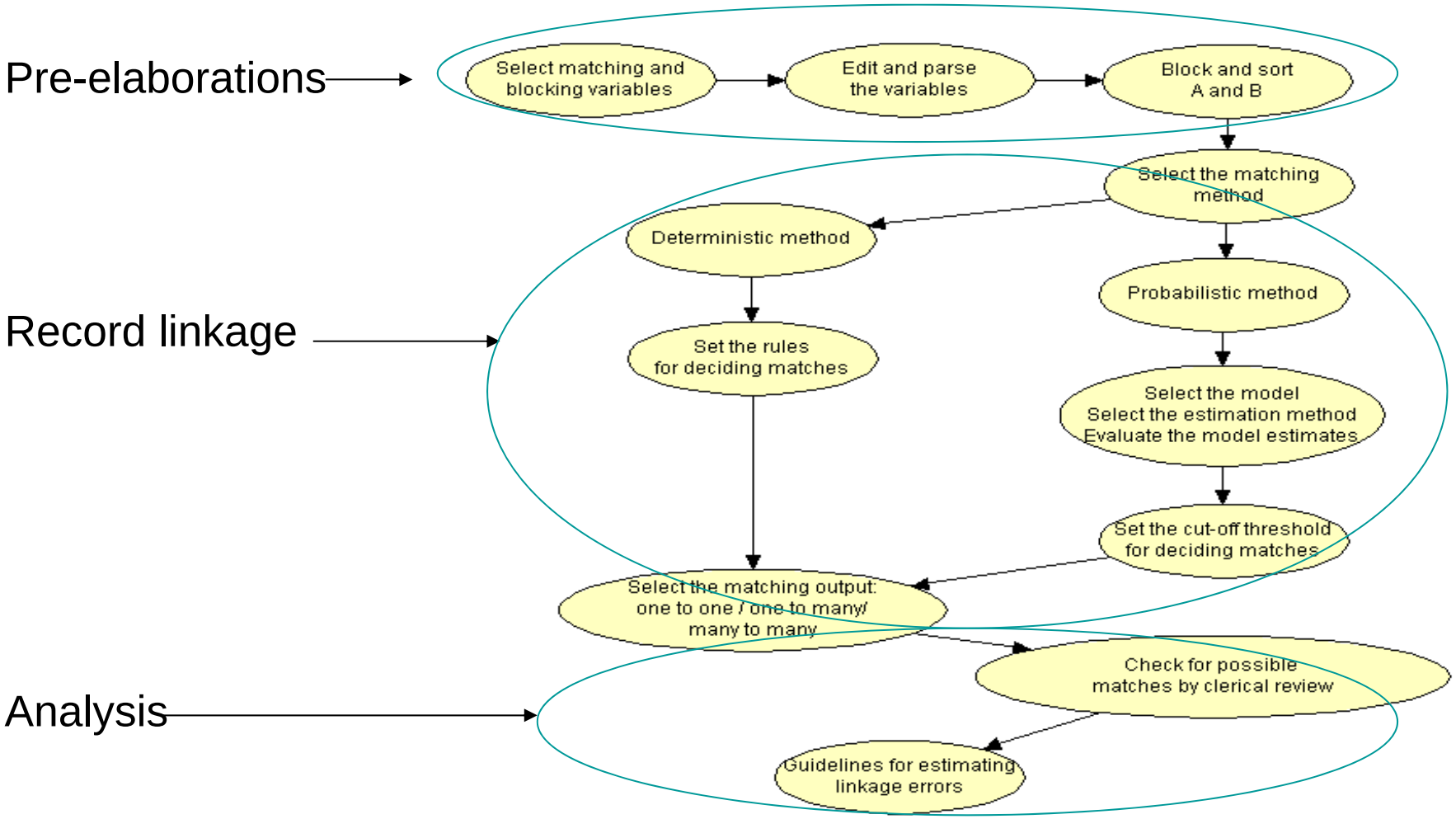
The record linkage phases

1) Pre-elaborations
2) Record linkage
3) Analysis

An ONS report (Gill et al, 2001) considers the following graph as a description of the record linkage phases


Select matching and blocking variables
Edit and parse the variables
Block and sort A and B
Select the matching method
Deterministic method
Probabilistic method
Set the rules for deciding matches
Select the model
Select the estimation method
Evaluate the model estimates
Set the cut-off threshold for deciding matches
Select the matching output: one to one / one to many/ many to many
Check for possible matches by clerical review
Guidelines for estimating linkage errors

The record linkage phases



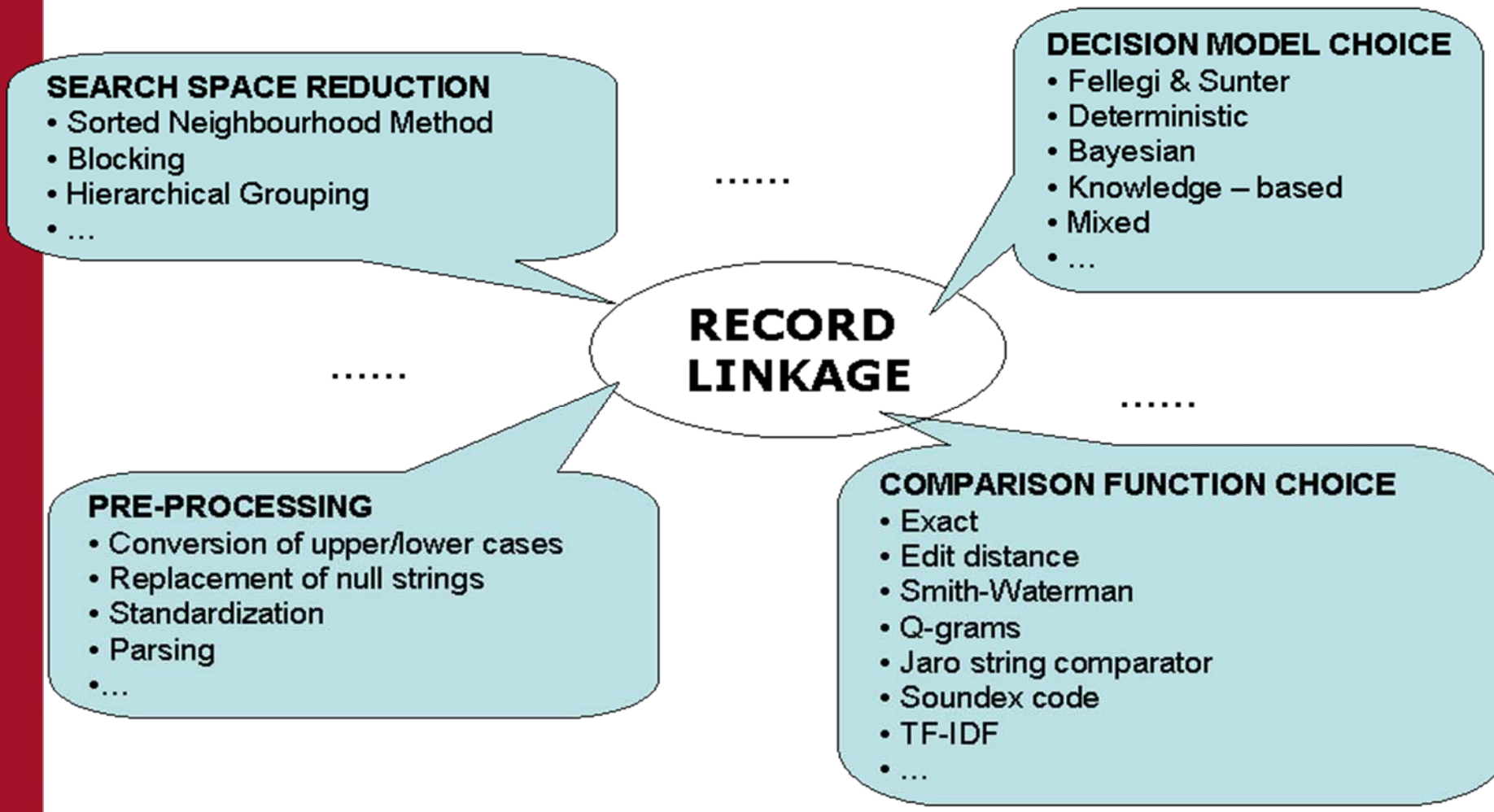
Decompose RL in phases

- Decompose a RL project in its constituting phases
- The best solution for all cases does not exist : choose the most appropriate technique for each phase, depending on application and data requirements, not only on practitioner's skill
- Dynamically build ad-hoc workflow for each RL problem

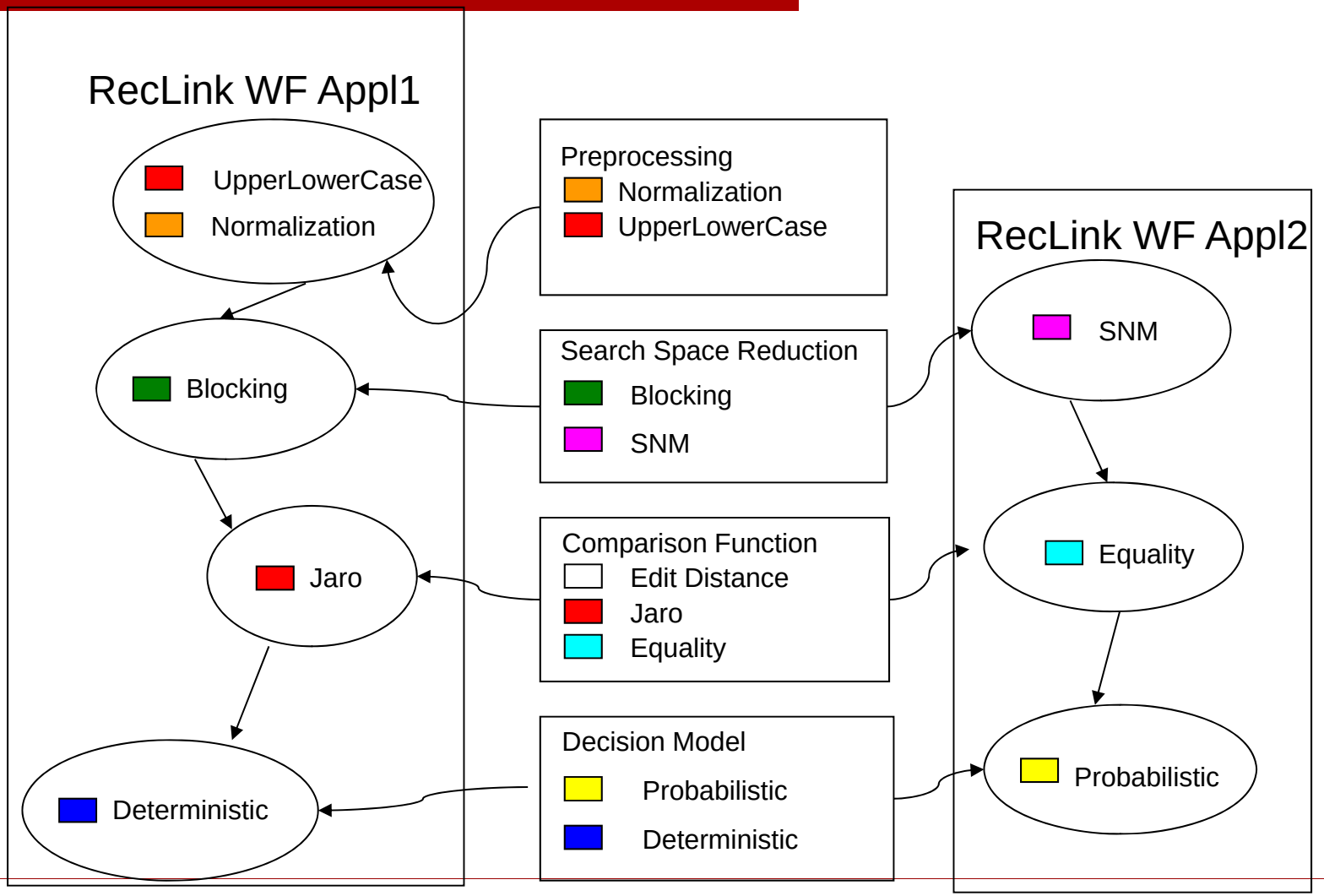
Decompose RL in phases

1. Pre-processing of the input files
2. Creation-Reduction of the search space of link candidate pairs
3. Choice of the matching variables
4. Choice of the comparison function
5. Choice of the decision model
6. Identification of unique links
7. RL evaluation

Choose the most appropriate techniques



Build ad-hoc RL workflows



Phases and techniques

5 november 10:50-13:00

Pre-processing of the input files

Creation-Reduction of the search space of candidate pairs

Choice of the matching variables

Choice of the comparison function

Identification of unique links

6 november 9:00-10:30

Choice of the decision model

6 november 10:50-13:00

RL evaluation

Statistical methods for integration

In what sense can record linkage be interpreted as a statistical procedure?

Probabilistic record linkage is a statistical method because it is necessary to include some statistical steps in the procedure: estimation and test

Statistical phases of record linkage

Record linkage has two fundamental statistical phases:

1) A decision procedure
2) An estimation procedure of the elements necessary for the application of the decision rule

Furthermore, there are other important statistical phases, connected with the previous two

3) Definition of the accuracy of the record linkage procedure and the detection of the probability of false match
4) Analysis of the linked data set

Bibliography

Belin, Rubin (1995) A method for calibrating false-match rates in record linkage. JASA, 694-707

Fellegi (1997) “Record Linkage and Public Policy: A Dynamic Evolution”. In Alvey, Jamerson (eds) *Record Linkage Techniques*, Proceedings of an international workshop and exposition, Arlington (USA) 20-21 March 1997

Gill et al (2001) Methods for automatic record matching and linkage and their use in National Statistics. ONS methodological series no. 25

Gu L., Baxter R., Vickers D., Rainsford C., (2003); Record linkage: Current practice and future directions, Technical Report 03/83, CSIRO, Mathematical and Information Sciences. http://citeseer.ist.psu.edu/585659.html

Herzog, Scheuren, Winkler (2007) Data quality and record linkage. Springer

Newcombe, Kennedy, Axford, James (1959) Automatic linkage of vital records. Science, 954-959

Statistics New Zealand (2006) Data integration manual