ESSnet Implementing Shared Statistical Services
Grant Agreement 566-831733-2018-FR-ESSnet-SERV2

http://www.cros-portal.eu/

Grant Agreement Number **566 - 831733-2018-FR-ESSnet-SERV2**

I3S - Implementing Shared Statistical Services

DeliverableD2.1.2

Relais service architecture

Authors: F. Amato, M. Bruno, P. Francescangeli,
G. Ruocco, L. Tosco, L. Valentino

Organization: Istat

Version: 1.0

Summary

1. Introduction

The Essnet Serv2 is a project launched in 2018 to develop new statistical services, either from scratch or from existing tool. The software Relais (Record Linkage At IStat) was selected for the relevance of data integration in the statistical process.

The first version of Relais dates back to 2007. The toolkit was designed as a software running on desktop computer. It is developed in Java and runs R programs to perform different methods for record linkage. Data are stored in a MySQL database.

This document provides: i) the description of Relais current version (AS-IS) that is already composed by self-contained modules; ii) an overview of the software re-engineering, adopting a generalized architectural pattern (Relais in IS2).

The following analysis is focused on the business and application layers. The assessment of the technology layer is still in progress.

2. Current architecture

Relais (Record Linkag At IStat) is a software designed and developed with the aim of providing an integrated environment to solve record linkage problems based on the belief that a record linkage process is structured in sub-phases that can be combined together in order to obtain the best process for the data being processed. Figure 1 shows an overview of the AS-IS architecture from the business layer point of view.

As shown in Figure 1, the current realization of Relais offers, for each process step, many business services that realize different techniques to solve the steps. From the application layer point of view, Relais implements each business service with a specific Java module that, in two cases, calls an R program. Currently, Relais is a desktop module with a graphic user interface that, step-by-step, guides the user in the realization of the record linkage process. Figure 1 shows in blue the application modules callable to run the chosen solution step while in dark blue are shown the most important data structures (mySql tables) used by the software.

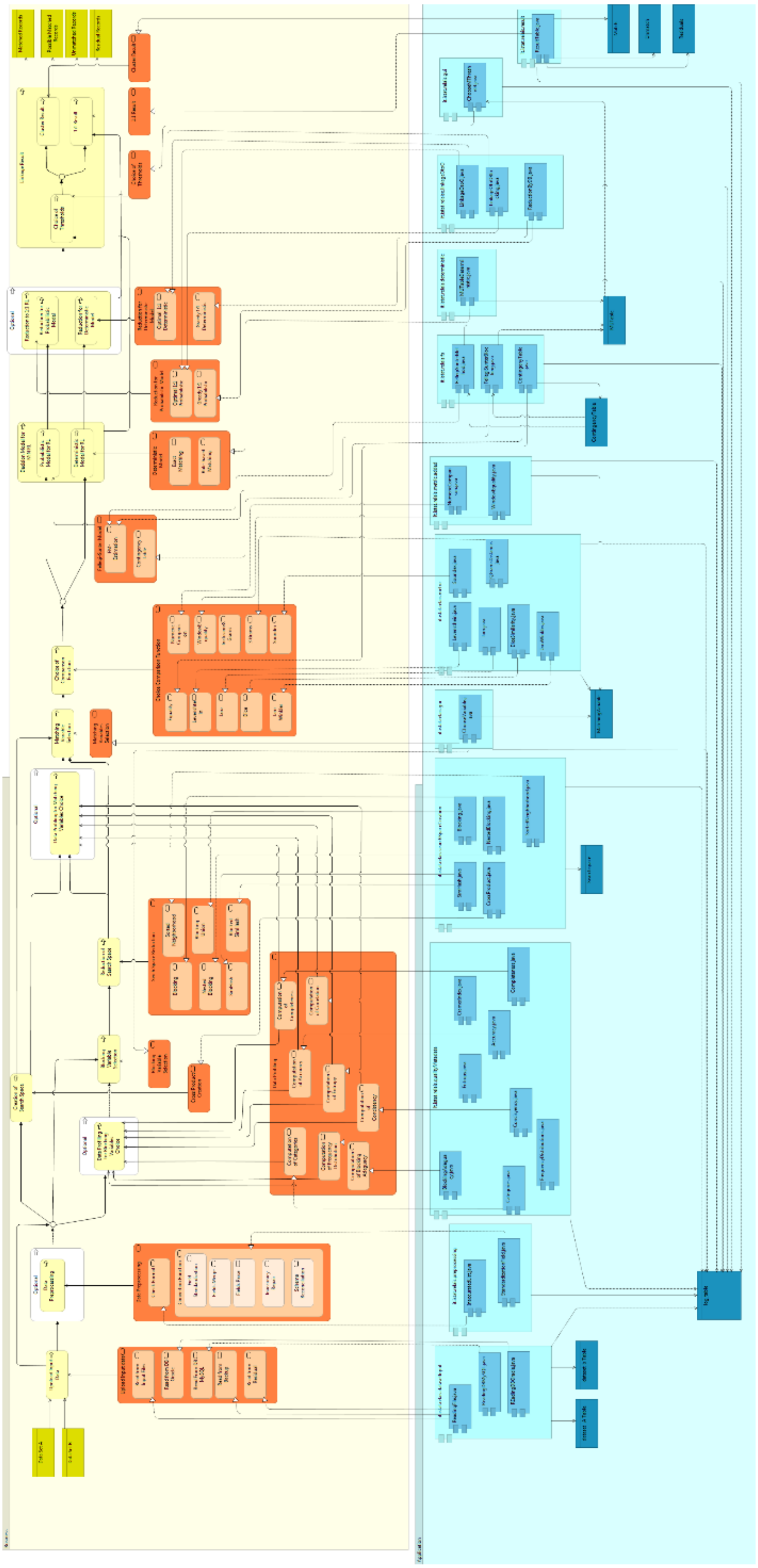


Figure 1. Relais AS-IS architecture

2.1 Business layer

The designed business layer shows that a record linkage process can be realized by different workflows composed by different steps, depending from the data to be linked. Different pattern can be composed allowing mainly a deterministic, or a probabilistic record linkage approach. Some process steps are mandatory while other steps are optional.

The core steps for realizing a record linkage process are:

1. Upload input data: different input data format are admitted (csv file, mySql Tables, Oracle Tables);

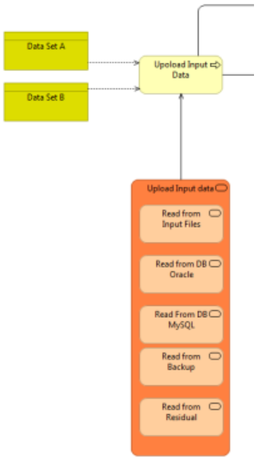


Figure 2. Upload input data business process and service

2. Search Space creation: it can be realized simply via a cross product or a search space reduction step can be performed. In this second case, different methods can be performed, namely: Blocking, Nested Blocking, Sorted Neighborhood, Blocking Union, Simhash and Blocked Simhash. If a space reduction step is performed, a previous step of blocking variable selection must be performed.

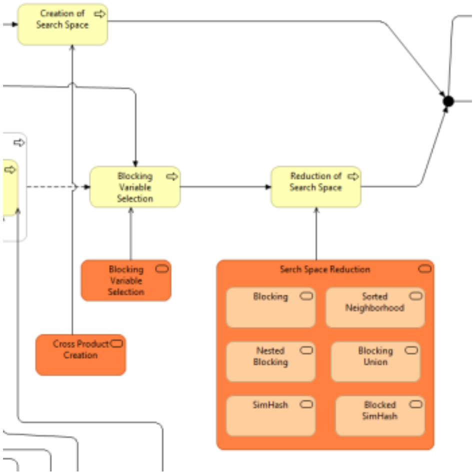


Figure 3. Search Space Creation business process and service

3. Matching variable and comparison function selection: In this step, the best matching variable must be selected and for each of theme a comparison function must be associated. These variables and functions are used to perform the following step.

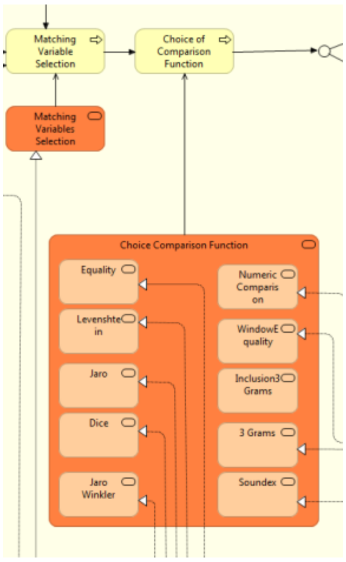

Matching Variable Selection
Choice of Comparison Function
Matching Variables Selection
Choice Comparison Function
Equality
Numeric Comparis on
Levenshte in
WindowE quality
Jaro
Inclusion3 Grams
Dice
3 Grams
Jaro Winkler
Soundex

Figure 4. Matching variable and comparison function selection business process and service

4. Decision Model: In this step, a determinist or probabilistic record linkage model can be applied. In the first case, a further choice can be made between rule based or exact matching. The result of this step is a M:N record linkage, that is a record of the first data source is linked to N records of the second data source and a record of the second data source is linked to M records of the first data source.

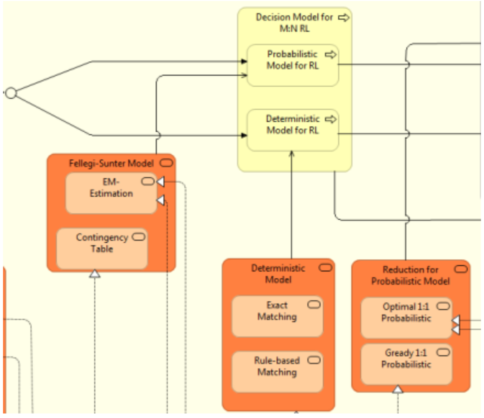

Decision Model for M:N RL
Probabilistic Model for RL
Deterministic Model for RL
Fellegi-Sunter Model
EM-Estimation
Contingency Table
Deterministic Model
Exact Matching
Rule-based Matching
Reduction for Probabilistic Model
Optimal 1:1 Probabilistic
Gready 1:1 Probabilistic

Figure 5. Decision model business process and service

5. Linkage Result: this step is composed of two sub-steps namely choice of thresholds and results creation. Firstly, two thresholds must be chosen: the match threshold and the un-match threshold. These thresholds discriminates couples of records defining three sets: Match, Unmatch and Possible Match records. The first set contains couple of records that surely relates to the same real world entity; the second set contains couple of records that surely do not refer to the same real world entity. The third set contains couple of records for which a human revision is needed to discriminate whether a couple refer to the same entity or not.

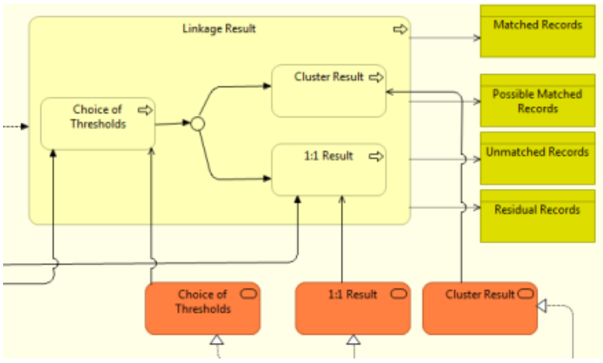


Figure 6. Linkage result business process and service

In the process creation, are available four optional process steps:

- Preprocessing: different “data manipulation” are realizable to clean and better format data to perform a better record linkage process.

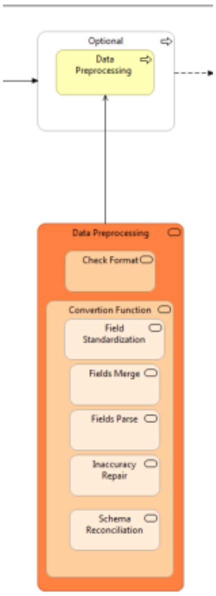


Figure 7. Preprocessing business process and service

- Data profiling for blocking variable choice. In this phase different indicators are available that help user to analyze variables and make better choice for blocking phase.

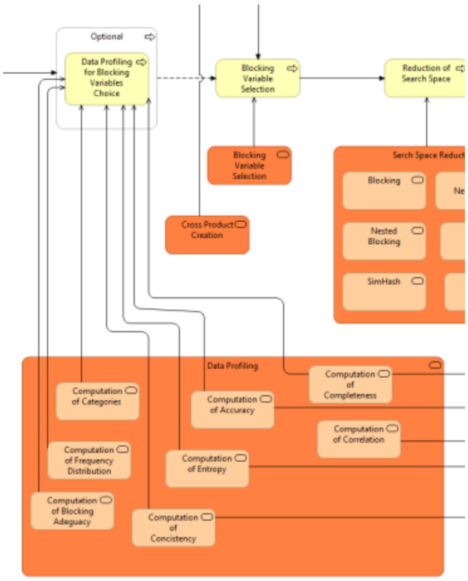

Optional
Data Profiling for Blocking Variables Choice
Blocking Variable Selection
Reduction of Search Space
Blocking Variable Selection
Serch Space Reduct
Blocking
Ne
Nested Blocking
SimHash
Cross Product Creation
Data Profiling
Computation of Categories
Computation of Completeness
Computation of Accuracy
Computation of Correlation
Computation of Frequency Distribution
Computation of Entropy
Computation of Blocking Adeguacy
Computation of Concistency

Figure 8. Data profiling for blocking variable choice business process and service

- Data profiling for matching variable choice. In this phase different indicators are available that help user to analyze variables and make better choice for matching records.

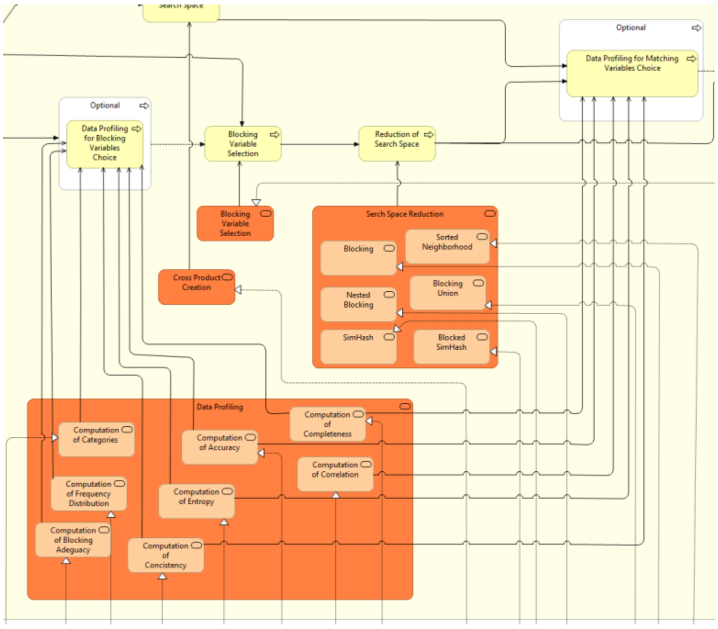

Search Space
Optional
Data Profiling for Matching Variables Choice
Optional
Data Profiling for Blocking Variables Choice
Blocking Variable Selection
Reduction of Search Space
Blocking Variable Selection
Serch Space Reduction
Blocking
Sorted Neighborhood
Nested Blocking
Blocking Union
SimHash
Blocked SimHash
Cross Product Creation
Data Profiling
Computation of Categories
Computation of Completeness
Computation of Accuracy
Computation of Correlation
Computation of Frequency Distribution
Computation of Entropy
Computation of Blocking Adeguacy
Computation of Concistency

Figure 9. Data profiling for matching variable choice business process and service

- Reduction to 1:1 record linkage: as previously seen, the decision models give as result a N:M linkage. This result is not always the best one; a unique correspondence between records can be desired. In this case, the user can choose between two different methods: optimal solution or greedy solution.

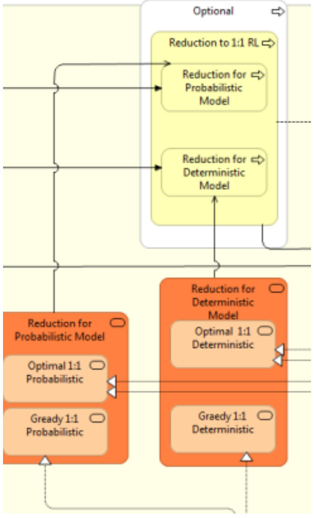

Optional
Reduction to 1:1 RL
Reduction for Probabilistic Model
Reduction for Deterministic Model
Reduction for Deterministic Model
Optimal 1:1 Deterministic
Gready 1:1 Deterministic
Reduction for Probabilistic Model
Optimal 1:1 Probabilistic
Gready 1:1 Probabilistic

Figure 10. Reduction to 1:1 record linkage business process and service

2.2 Application layer

Currently, the application layer consist of java modules realizing the business services. Figure 1 shows the main modules that realize the services. A graphical interface helps the user to interact with data and helps him to design and to run the record linkage process.

Referring to the core business steps, described in the previous section, in the following we report the main java module called from the graphical interface to execute the business service.

1. Upload input data: as shown in the following excerpt of Figure 1, the main java modules executing the upload business service are ReadingFile.java, ReadingDBMySQL.java and ReadingDBOracle.java, depending from the input format.

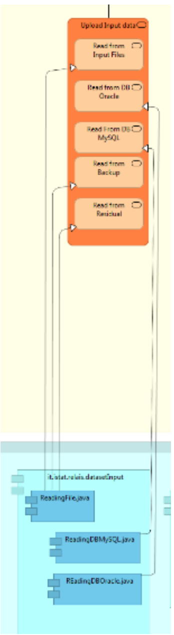


Figure 11. Upload input data application module

2. Search Space creation: as shown in the following excerpt of Figure 1, the Cross Product Creation service is realized by the CrossProduct.java java module. The SimHash.java, Blocking.java, NestedBlocking.java and SortedNeighborhood.java java modules realize the corresponding business service of the Search Space Reduction business service.

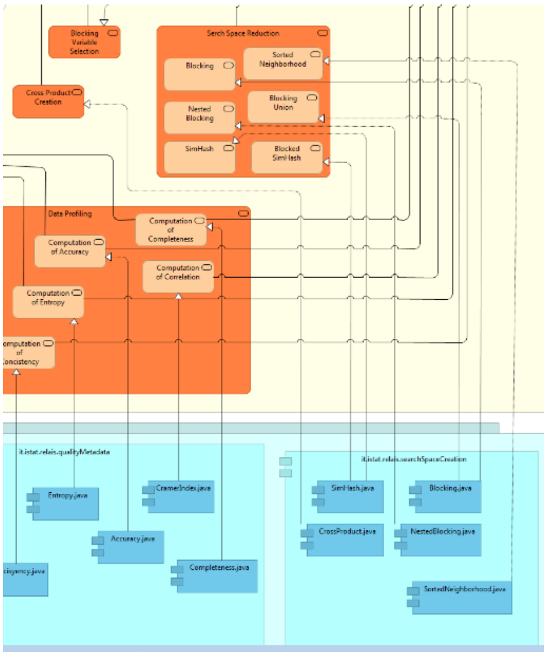


Figure 12. Search Space creation application module

3. Matching variable and comparison function selection: as shown in the following excerpt of Figure 1, the Matching Variable Selection service is realized by the java module ChooseVariable.java, while each comparison function is realized by the corresponding java module, specifically: Levenstein.java implements Levenstein service, Jaro.java implents Jaro service, and so on.

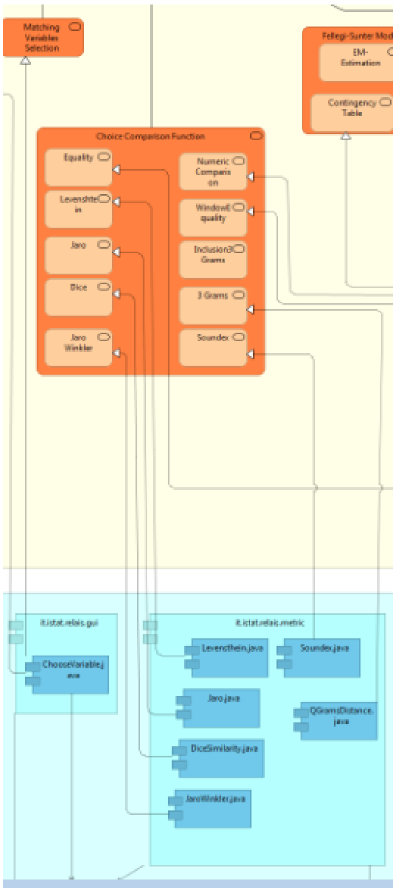


Figure 13. Matching variable and comparison function selection apllication module

4. Decision Model: two java modules implement the probabilistic method, realized by the Fellegi-Sunter Method business service namely: Contingencytable.java and FellegiSunterMethod.java (that invoke an R program). The MUTableDeterminist.java java module implements the deterministic method business service.

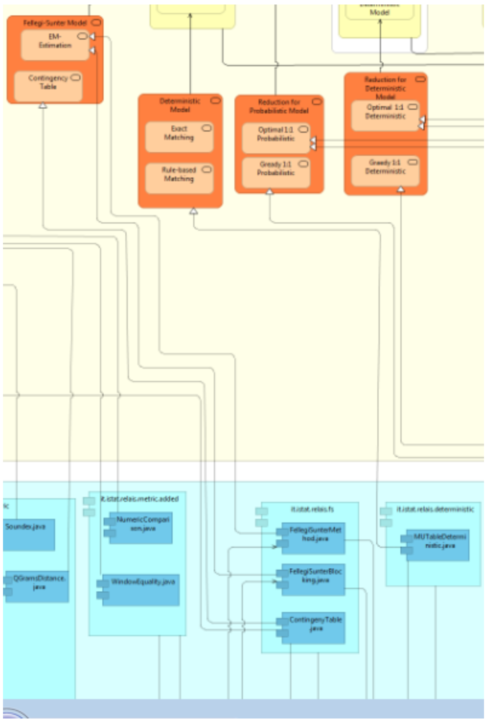


Figure 14. Decision Model application module

5. Linkage Result: the Choice Of Thresholds business service is implemented by the ChooseMThreshold.java java module. The cluster result business service is implemented by the ResultTable.java java module.

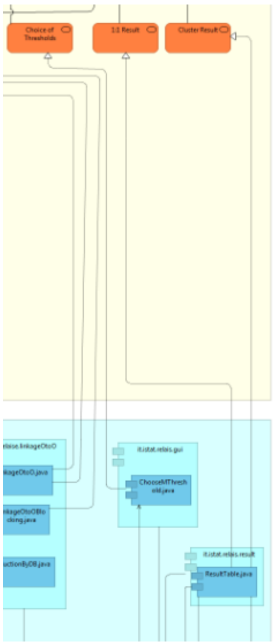


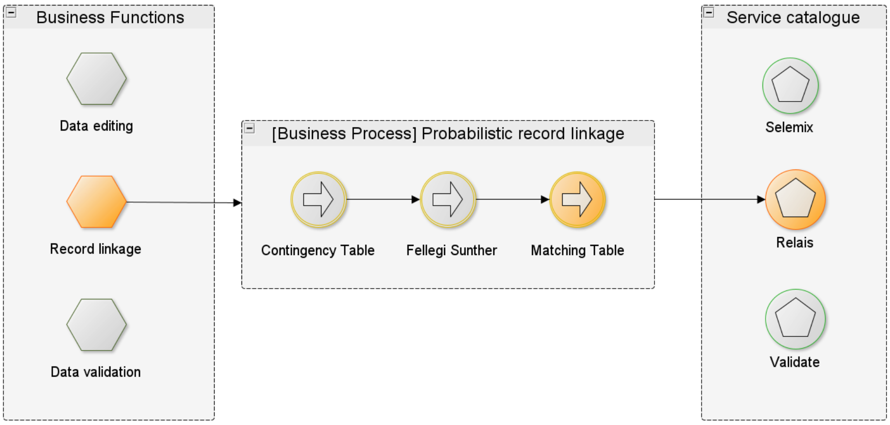
Figure 15. Linkage result application module

3. Statistical Service Generalized Architecture (IS^2)

Istat has realized a generalized environment (Istat Statistical Service - IS^2) that allows to select statistical services from a catalogue and execute them through a web application (IS^2 workbench).
The IS^2 Workbench has been designed to offer a set of functionalities that allow to:

1. **Select a business function:** the landing page contains the list of available business functions classified according to GSBPM phases (e.g. ReLais performs GSBPM 5.1 "Data integration"). A business function is a high level goal (What) that can be realized by one or more statistical processes, implemented by one or more services available in the catalogue.
2. **Select a business process:** the system provides the list of available processes for the selected function (e.g. Probabilistic Record Linkage or Deterministic Record Linkage). A business process is implemented by a set of process steps. Each process step is linked to a statistical service available in the catalogue. Statistical services perform specific statistical method, implemented in an open source language.

   The following figure shows the link between the "Record linkage" business function, available in the workbench, the "Probabilistic record linkage" process and the related statistical service (Relais).



3. **Upload process input data:** in order to launch a process, the system requires the specification of input data to process. The initial set of data may include a list of rules, and/or other parameters used by the statistical method embedded in the process steps.
4. **Set process metadata:** a statistical service may require further information, depending on the statistical function to perform. This set of metadata is provided by the user and is usually tied to input data structure, or concerns model parameters (e.g. specification of matching variables in the datasets to be linked, setting of matching/unmatching thresholds).
5. **Execute a business process according to a predetermined workflow:** this function allows to execute the process previously configured. The main advantages of this approach are:
   - orchestrate and control the process steps execution;
   - document data transformation;
   - enhance process standardization and reduce software overlapping.

6. **Analyze output data:** each statistical service will offer a set of functionalities to help the user to assess the output of each step, and/or the final result. Depending on user requirements, output analysis will be performed at micro and/or macro level (aggregated reports).
7. **Auxiliary functionalities for data management:** regardless the statistical service selected, the user can perform data sorting, data visualization, data filtering and data export in csv format.

3.1 Architecture components

An overview of the main architectural components of IS^2 framework is shown in the figure below.

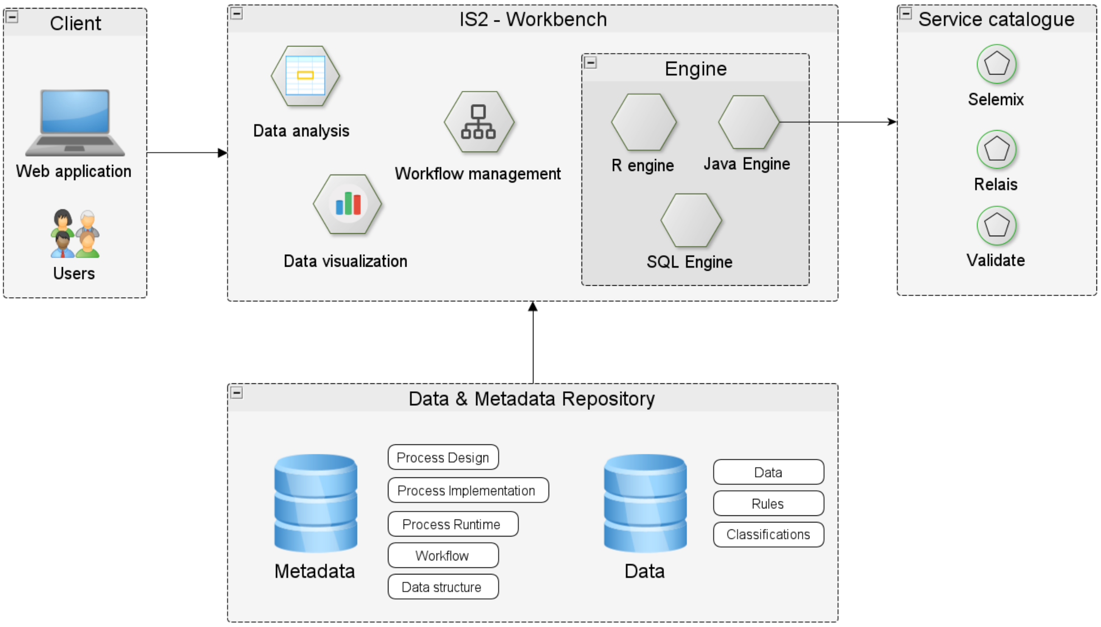


Figure 16. IS^2 Architecture

The main components of IS^2 architecture can be grouped in the following subsets:

1) **Service catalogue:** set of statistical services implemented. A statistical service is a software component performing a statistical method that can be invoked using REST protocol.

2) **Data/Metadata Repository:** the metadata model allows to execute a procedure regardless of input data structure. Metadata are grouped in the following interconnected subsets, based on GSIM concepts:

- **Process Design:** the concepts (Business Function, Business Process, Process Step, Business Service) defined in this section allow to model a statistical process at business level (What);
- **Process Implementation:** this section contains the list of libraries/packages executing the process steps and the statistical methods (How). For each library/package the metadata model allows to specify input/output and parameters;
- **Process Runtime**: describes the runtime execution in terms of data input, output, parameters and function signature (When);
- **Data structure**: describes the initial data and specify the core variables involved in data process. This subset contains also the classifications and/or the auxiliary data sources. By mapping initial data with

standardized metadata, input data are transformed in working data, that will be processed by the statistical services.

- **Workflow management:** this section contains the concepts (Workflow, Ruleset, Rule) that allow to describe and manage the sequence of process steps through a set of Rules.

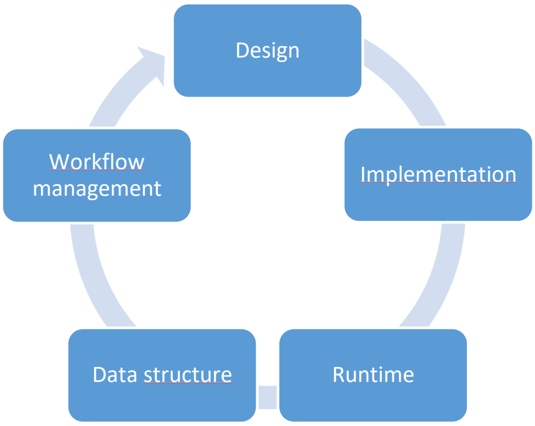


Figure 17- Metadata repository main concepts

A detailed description of IS^2 database, containing data and metadata, is provided in Annex 2.

3) **IS^2 Workbench**, composed by the following set of components:

- **Engines:** these modules allow to invoke the statistical services available in the catalogue, implemented in several programming languages (R, Java and SQL).
- **Work session management**: the workbench provides a set of functionalities to manage user working sessions. More precisely:
    1. **Work session:** this step includes data upload and preprocessing. In addition to the information to process, input data may include auxiliary information split in different datasets.
    2. **Data processing:** in this step, the user can classify and manage the information to process by: i) assigning specific roles to some variables (e.g. identification variable, classification, core variables); ii) selecting auxiliary information (if needed); iii) setting the model parameters.
- **Workflow management:** a statistical service, regardless of the implemented method, is supposed to execute the following tasks: i) upload and manage input data, to provide initial data to process; ii) set parameters and variables; iii) run method; iv) analyze output.
- **Auxiliary Components:** a statistician may need to perform basic statistics, or create new variables by transforming the existing ones, or select a subset of records and/or variables. These components perform data analysis and data visualization.

4. Reengineering Relais in IS^2

We decided to reengineer Relais starting from the business process, shown in Figure 18, that realizes a probabilistic record linkage workflow applying the Fellegi-Sunter method. The process is composed by the following steps:

1. Selection of datasets for record linkage.
2. Creation of search space using the cross-product method.
3. Matching variable and comparison function selection.
4. Application of the decision model that implements the Fellegi-Sunter probabilistic method.
5. Creation of the resulting datasets (match, unmatch, possible match).

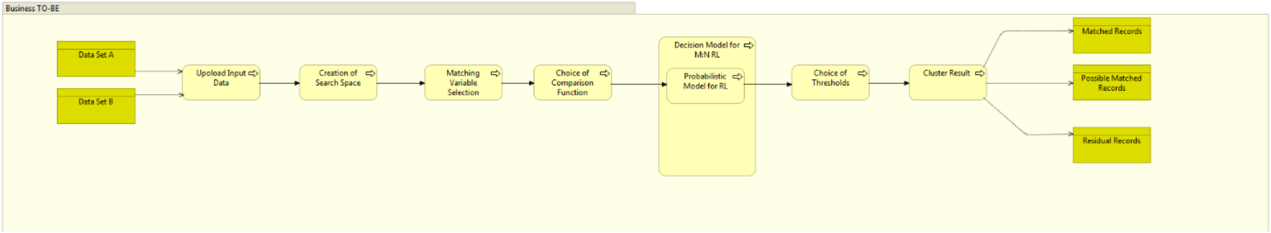


Figure 18. Probabilistic record linkage workflow

To implement "Probabilistic record linkage workflow" in IS^2 environment, the process steps have been reviewed as shown in Figure 19.

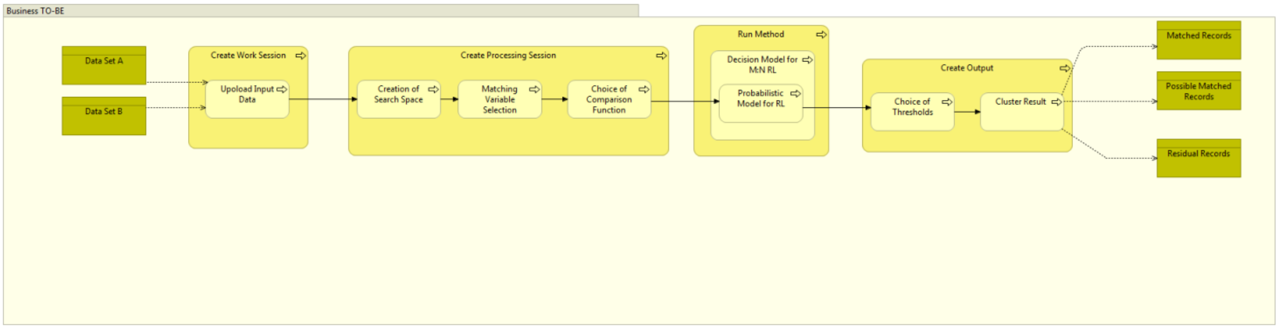


Figure 19. Relais business workflow implemented first, compliant to the IS2 business architecture

The following table shows the relation between the AS-IS process steps and the new IS^2 compliant ones.

AS-IS Relais process steps	IS^2 Relais process steps
Upload Input Data	Create Working Session – upload input data
Creation of Search Space	Create Working Session – select method
Matching variable Selection	Create Processing Session – specify core variables
Choice of Comparison Function	Create Processing Session – set parameters
Probabilistic model for RL	Run method
Choice of thresholds	Create Processing Session – set parameters
Cluster result	Run Method – create output Analyse output

Table 1. Correspondence between old and new process steps

Annex 1. Example of inputs and outputs of the business steps implemented in Relais-IS2.

The following example aims to clarify the contents of the inputs and outputs of the business steps composing the business function of Relais currently implemented in IS2 (as detailed in Section **Errore. L'origine riferimento non è stata trovata.**) .

First input file (Data set A):

Id_DSA	Name	Surname	Address	Phone Number
1	Alessandro	Gassmann	v. Roma,11	333333333
2	Marcello	Mastroianni	Viale Europa 16	333222222
3	Pasquale	Banfi	Via Mar Tirreno 1	333111111
4	Francesco	Zalone	Viale Adriatico 11	333000000

Second input file (Data set B):

Id_DSB	Name	Surname	Address	Place of Birth	Date of Birth
1	Lino	Banfi	Via Mar Tirreno 1	Andria	09/07/1936
2	Diego	Abatantuono	Via Roma 128	Milano	20/05/1955
3	Alessandro	Gassman	v. Roma,11	Roma	24/02/1965
4	Pierfrancesco	Favino	Via torino 16	Roma	24/08/1969
5	Pasquale	Zalone	Viale Adriatico 114	Bari	03/06/1977
6	Marcello	Mastroianni	Viale Europa	Fontana Liri	28/09/1924

Starting from the input files, the search space business step, evaluated applying the cross-product method gives the following the resulting table:

ID_DSA	ID_DSB
A_1	B_1
A_1	B_2
A_1	B_3
A_1	B_4
A_1	B_5
A_1	B_6
A_2	B_1
A_2	B_2
A_2	B_3
A_2	B_4
A_2	B_5
A_2	B_6
A_3	B_1
A_3	B_2
A_3	B_3
A_3	B_4
A_3	B_5

A_3	B_6
A_4	B_1
A_4	B_2
A_4	B_3
A_4	B_4
A_4	B_5
A_4	B_6

Choosing Name, Surname and Address as matching variables and choosing for all of them Jaro as comparison function with a threshold equal to 0.8, we obtain the following contingency table:

Name	Surname	Address	Frequency
0	0	0	19
0	0	0	0
0	1	0	0
0	1	1	2
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	2

Assuming that applying the Fellegi-Sunter probabilistic model, we obtain the following result:

Name	Surname	Address	Frequency	Probability
0	0	0	19	0.1
0	0	1	0	0.2
0	1	0	0	0.2
0	1	1	2	0.9
1	0	0	1	0.5
1	0	1	0	0.2
1	1	0	0	0.2
1	1	1	2	1

To each couple of records of the search space table corresponds a probability, as shown the following table:

ID_DSA	Name_DSA	Surname_DSA	Address_DSA	ID_DSB	Name_DSB	Surname_DSB	Address_DSB	Pattern	Prob
A_1	Alessandro	Gassmann	v. Roma,11	B_1	Lino	Banfi	Via Mar Tirreno 1	000	0.1
A_1	Alessandro	Gassmann	v. Roma,11	B_2	Diego	Abatantuono	Via Roma 128	000	0.1
A_1	Alessandro	Gassmann	v. Roma,11	B_3	Alessandro	Gassman	v. Roma,11	111	1
A_1	Alessandro	Gassmann	v. Roma,11	B_4	Pierfrancesco	Favino	Via torino 16	000	0.1
A_1	Alessandro	Gassmann	v. Roma,11	B_5	Pasquale	Zalone	Viale Adriatico 114	000	0.1
A_1	Alessandro	Gassmann	v. Roma,11	B_6	Marcello	Mastroianni	Viale Europa	000	0.1

A_2	Marcello	Mastroianni	Viale Europa 16	B_1	Lino	Banfi	Via Mar Tirreno 1	000	0.1
A_2	Marcello	Mastroianni	Viale Europa 16	B_2	Diego	Abatantuono	Via Roma 128	000	0.1
A_2	Marcello	Mastroianni	Viale Europa 16	B_3	Alessandro	Gassman	v. Roma,11	000	0.1
A_2	Marcello	Mastroianni	Viale Europa 16	B_4	Pierfrancesco	Favino	Via torino 16	000	0.1
A_2	Marcello	Mastroianni	Viale Europa 16	B_5	Pasquale	Zalone	Viale Adriatico 114	000	0.1
A_2	Marcello	Mastroianni	Viale Europa 16	B_6	Marcello	Mastroianni	Viale Europa	111	1
A_3	Pasquale	Banfi	Via Mar Tirreno 1	B_1	Lino	Banfi	Via Mar Tirreno 1	011	0.9
A_3	Pasquale	Banfi	Via Mar Tirreno 1	B_2	Diego	Abatantuono	Via Roma 128	000	0.1
A_3	Pasquale	Banfi	Via Mar Tirreno 1	B_3	Alessandro	Gassman	v. Roma,11	000	0.1
A_3	Pasquale	Banfi	Via Mar Tirreno 1	B_4	Pierfrancesco	Favino	Via torino 16	000	0.1
A_3	Pasquale	Banfi	Via Mar Tirreno 1	B_5	Pasquale	Zalone	Viale Adriatico 114	100	0.5
A_3	Pasquale	Banfi	Via Mar Tirreno 1	B_6	Marcello	Mastroianni	Viale Europa	000	0.1
A_4	Francesco	Zalone	Viale Adriatico 11	B_1	Lino	Banfi	Via Mar Tirreno 1	000	0.1
A_4	Francesco	Zalone	Viale Adriatico 11	B_2	Diego	Abatantuono	Via Roma 128	000	0.1
A_4	Francesco	Zalone	Viale Adriatico 11	B_3	Alessandro	Gassman	v. Roma,11	000	0.1
A_4	Francesco	Zalone	Viale Adriatico 11	B_4	Pierfrancesco	Favino	Via torino 16	000	0.1
A_4	Francesco	Zalone	Viale Adriatico 11	B_5	Pasquale	Zalone	Viale Adriatico 114	011	0.9
A_4	Francesco	Zalone	Viale Adriatico 11	B_6	Marcello	Mastroianni	Viale Europa	000	0.1

Choosing 0.8 as match threshold and 0.4 as unmatch threshold, we obtain the following results.

File of Match record:

Id_DSA	Name	Surname	Address	Phone Number	Id_DSB	Name	Surname	Address	Place of Birth	Date of Birth
A_1	Alessandro	Gassmann	v. Roma, 11	333333333	B_3	Alessandro	Gassman	v. Roma, 11	Roma	24/02/1965
A_2	Marcello	Mastroianni	Viale Europa 16	333222222	B_6	Marcello	Mastroianni	Viale Europa	Fontana Liri	28/09/1924
A_3	Pasquale	Banfi	Via Mar Tirreno 1	333111111	B_1	Lino	Banfi	Via Mar Tirreno 1	Andria	09/07/1936
A_4	Francesco	Zalone	Viale Adriatico 11	333000000	B_5	Pasquale	Zalone	Viale Adriatico 114	Bari	03/06/1977

File of Possible Match record:

Id_DSA	Name	Surname	Address	Phone Number	Id_DSB	Name	Surname	Address	Place of Birth	Date of Birth
A_3	Pasquale	Banfi	Via Mar Tirreno 1	333111111	B_5	Pasquale	Zalone	Viale Adriatico 114	Bari	03/06/1977

File of first file residuals:

Id_DSA	Name	Surname	Address	Phone Number

File of second file residuals:

Id_DSB	Name	Surname	Address	Place of Birth	Date of Birth
2	Diego	Abatantuono	Via Roma 128	Milano	20/05/1955
4	Pierfrancesco	Favino	Via torino 16	Roma	24/08/1969

Annex 2. Data base schema

In this section we describe the database implemented for the workbench. The database has been designed according to a process-oriented approach. The model is based on the following concepts:

- **Process Design:** the concepts (Business Function, Business Process, Process Step, Business Service) defined in this section allow to model a statistical process at business level (What);
- **Process Implementation:** this section contains the list of libraries/packages executing the process steps and the statistical methods (How). For each library/package the metadata model allows to specify input/output and parameters;
- **Process Runtime**: describes the runtime execution in terms of data input, output, parameters and function signature (When);
- **Data structure**: describes the initial data and specify the core variables involved in data process. This subset contains also the classifications and/or the auxiliary data sources. By mapping initial data with standardized metadata, input data are transformed in working data, that will be processed by the statistical services.
- **Workflow management:** this section contains the concepts (Workflow, Ruleset, Rule) that allow to describe and manage the sequence of process steps through a set of Rules.

The different subsets are shown in Figure 20.

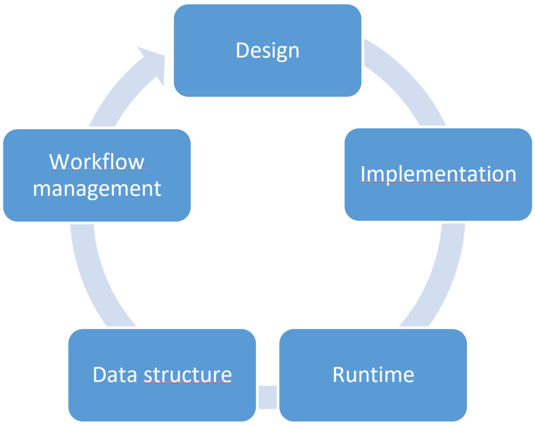


Figure 20 - Database concepts

The script to generate the database is available on github:

https://github.com/mecdcme/is2

The ‘db’ folder contains the script is2.sql.

A brief description of the core tables of each subset will be provided in the following sections.