Record Linkage: Purposes, Practices and Quality Issues

Heidi Koumarianos[*], *Olivier Lefebvre*[**] *and Lucas Malherbe*[***]

Record linkage reconciles individual data that are taken from various data files. It is used more and more frequently in official statistics, whether for analytical purposes, to investigate new subjects, or to improve production processes. Such statistical processing often raises issues relating to the imperfections of the data used and their volume.

Given the issues related to the compliance with data confidentiality, especially regarding personal data protection, an appropriate legal framework is required to implement them. A good knowledge of the data used and particular care in determining the parameters are also necessary to ensure the best possible quality of the results, since record linkage is never 100% reliable.

The quality of these matches is therefore a major issue for official statistics, and requires a direct assessment of the process, necessarily supplemented by a study of the statistical impact of the linkage on the data produced.

* Methodologist, *Direction de la méthodologie et de la coordination statistique et internationale* (Methodology, Statistical Coordination and International Relations Directorate, DMCSI), INSEE.
heidi.koumarianos@insee.fr

** Head of the RÉSIL Program, INSEE.
olivier.lefebvre@insee.fr

*** At the time of writing, data scientist, DMCSI, INSEE.

The measurement of household income and the monitoring of the integration of graduates into professional life have one thing in common: they both rely on the use of multiple data sources, which must be combined with each other. At the most granular level, the data concerning each individual or household must be collected in each of the sources. This makes it possible to cover all sources of income for households, whether they are taxable or not, or to follow the career paths of graduates, and in particular the conditions they face when entering the labour market.

Combining different sources allows for richer and more efficient observation. Record linkage (the operation that allows sources to be linked and combined) can be used in many fields and is in some ways a collection technique, with inherent technical constraints, methodological challenges, legal framework and ethical issues. Most statistical offices use this technique for the production of statistical data, associated with the increasing use of administrative data, which are often very precise yet highly specific in terms of their content, and which therefore need to be supplemented.

INSEE and official statistics more generally have been performing data linking operations for many years now. This practice has gradually expanded thanks to the development of efficient data processing techniques and greater access to source files. Examples of this include Fidelimmo (André and Meslin, 2022), which makes it possible to more effectively analyse the property wealth of households and the redistributive impacts or lack thereof of property tax, InserJeunes (Midy, 2021) for measuring the integration into professional life of apprentices and SIRUS[1] (Hachid and Leclair, 2022), the backbone of business statistics.

▶ Linking to Enrich or Better Understand Sources...

Record linkage, when performed for statistical purposes, initially allows additional information to be provided to an existing statistical file. More generally, a distinction can be made between several different uses of linkage:

- **To supplement the coverage of the analysis**: the FILOSOFI[2] system, for example, uses various sources to reconstruct household income, whether it comes from employment or social benefits. It provides a more comprehensive view of household incomes, on a granular geographical scale, which can be as focused as the neighbourhoods of a city, if they are of sufficient size.

- **To provide clarity on certain phenomena**: for example, linking files of higher education graduates with employment files makes it possible to describe the integration into professional life of young graduates.

1 SIRUS: *Système d'immatriculation au répertoire des unités statistiques* (Statistical Business Register).
2 FILOSOFI: a set of indicators on located social and fiscal incomes.

• **To understand the impact of a social benefit or support for companies**: linking the file of beneficiaries with a file describing their situation (employment, success in higher education or financial outcomes) makes it possible to determine whether the support has generated any effects and to assess the impact of the benefit.

• **To better understand the content of the sources analysed**: for example, linking files of job seekers with those of the *enquête Emploi* (Labour Force Survey) has made it possible to better understand the various developments in unemployment as defined by the International Labour Office, and the number of job seekers registered with France Travail[3].

▶ … or to Improve Production Processes

In addition to the construction of new data, combining sources makes it possible to significantly improve statistical production processes. The phase in which information is gathered or the quality of sources is monitored or evaluated is modified (consistency with the concepts to be measured and coverage measurement). More precisely, record linkage makes it possible to:

• **Reduce survey questionnaires**: the principle is not to ask a household (or company) for information that it has already provided to an administrative body, in accordance with the "tell us once" principle. For example, linking the *enquête Emploi* with tax data reduces the number of income-related questions.

• **Update a register or repository** (RNIPP[4], SIRENE[5], REU[6], RÉSIL[7], SIRUS, etc.): new entities are added to the register (in which case it is essential to check that they have not already been included, to avoid duplicates) or some characteristics are updated (in which case it is essential to check that the correct observation is updated). The quality of the registers is essential for the quality of the statistical processes (Espinasse and Roux, 2022; Demotes-Mainard, 2019).

• **Analyse the coverage of a source by linking it to a repository**: this is a significant advance in the analysis of developments or in the processing of "collection gaps", as well as for non-respondents to a survey (which is currently possible in business statistics with SIRUS and which may be possible in demographic and social statistics with RÉSIL).

▶ What is it all about?

In practice, record linkage refers to the action of reconciling data, at the level of each observation unit, of two data files A and B, either to enrich one of the files with additional or updated variables or to create a new file containing all or part of the variables of each of the files ***(Figure 1***). Record linkage can be for administrative (e.g. a check on

3 France Travail: France Travail replaced Pôle Emploi in 2023
4 RNIPP: *Répertoire national d'identification des personnes physiques* (National Register for the Identification of Individuals).
5 SIRENE: *Système national d'identification et du répertoire des entreprises et de leurs établissements* (National Enterprise and Establishment Register Database).
6 REU: *Répertoire électoral unique* (Single electoral register).
7 RÉSIL: *Répertoire Statistique des Individus et des Logements* (Statistical Register of Individuals and Dwellings). See the paper by Olivier Lefebvre on RÉSIL in this issue.

Figure 1 - An Example of Linking to Study the Link Between Profession and Qualification Level

Table A

Identifier | Identity traits (linking variables) | Statistical variable of interest

ID_A	Surname	Forename	Date of birth	Profession
A1	RAFIK	Kélian	03/02/2001	Teacher
A2	GIROUT	Claudia	05/08/1993	Beautician
A3	REGARD	Esteban	23/11/1997	Forklift operator
A4	MOREAU	Lya	12/11/2002	Police officer
A5	PEREZ	Maria	27/06/1987	Sales manager
A6	LEFEBVRE	Olivier	27/06/1999	Statistician

?

Table B

Identifier | Identity traits (linking variables) | Statistical variable of interest

ID_B	Sur-name	Fore-name	Date of birth	Highest qualification
B1	LEFEVRE	Oscar	02/04/1975	Technology University Diploma (DUT)
B2	RAFIK	Kélian	03/02/2001	Master's
B3	REGAR	Esteban	23/11/1997	Baccalaureate
B4	GIROUT	Claudia	05/08/1993	BEP vocational qualification
B5	PEREZ	Marina	27/06/1987	Doctorate

Table of results

Id	Date of birth	Profession	Highest qualification
R1	03/02/2001	Teacher	Master's
R2	05/08/1993	Beautician	BEP vocational qualification
R3	23/11/1997	Forklift operator	Baccalaureate
R4	27/06/1987	Sales manager	Doctorate

Reading note: From the two files A and B, which contain the profession and the qualification respectively, the aim is to build a result file that contains both the profession and the qualification of people included in the two files. The process of building this file is explained in the rest of the paper.

> ***If the files have a common identifier that is of good quality, the technical operation is simple... It is then simply a case of establishing a "link" using this identifier.***

rights), operational (e.g. merging customer files) or statistical use (***Box 1***).

The rest of this paper focuses on how to perform linking operations and their statistical uses.

In order to make this reconciliation relevant, it is necessary to know which line[8] of file B corresponds to each line of file A, ensuring that they both refer to the same observation unit. If the files have a common identifier that is of good quality, the technical operation is simple: two lines with the same identifier naturally refer to the same unit of observation. It is

8 A line is a record representing an individual.

then simply a case of establishing a "link" using this identifier (it is nevertheless necessary to deal with all the other aspects: legal and ethical framework, analysis of the statistical quality of the result, etc.). If no such link is established, there are two possible approaches:

- individual identification: a common identifier is found by comparing these two files A and B with a third file, C, which is often larger and acts as a repository. This identification step therefore aims to successively seek out, for the observations in files A and B, the line of file C to which they correspond, and to make the "link" using the corresponding identifier[9];

- the comparison of pairs: the two files are compared directly, searching among all the possible pairs of observations which correspond to the same entity; for this, specific techniques are used, based either on the application of rules of successive decisions or on probabilistic models.

This is record linkage. There is also statistical matching, based on units belonging to specific strata, which will not be discussed here. The first of these concepts is known as "record linkage", while the second is referred to as "propensity score matching" (Rosenbaum and Rubin, 1983).

▶ Box 1. Record linkage, Enrichment, Interconnection, combination: All Synonyms?

An **enrichment** of the data in one file by another file entails finding information concerning a given individual in two different files and then generating a third file with the data thus collected.

Although linking refers only to the first phase of this operation (linking together two observations relating to the same entity), statisticians often refer to this technique as "**record linkage**" (or data linking). This is the term that appears in the Act of 1951 and the implementing decree of the Act for a Digital Republic*.

The authors of the 1978 Act on Information Technology, Data Files and Civil Liberties and the General Data Protection Regulation used the terms "**interconnection**", "**reconciliation**", or "**alignment**" of files. Interconnection and reconciliation of files are two forms of file alignment; the term interconnection is most often used for alignments with a high degree of automation, or even alignments that are fully automated.

Other terms may be used by statisticians to refer to linking: **combination or linkage** of files (this being the term used in Canada both by Statistics Canada and in the directive** that governs such operations).

To provide clarity regarding the subject and better establish the concept of record linkage in French law, the founding decree of RÉSIL provides a definition of record linkage***:

"These record linkage operations constitute alignments, as defined in Indent 3 of Section I of Article 33 of the Act [on Information Technology and Civil Liberties], between personal data recorded in the "*répertoire statistique des individus et des logements*" (Statistical Register of Individuals and Dwellings) and third-party statistical data sources. They result in the creation of new files, which constitute processing of personal data as defined in the [GDPR]."

* *Decree N° 2016-1930 of 28 December 2016 simplifying the advance formalities relating to processing for statistical or research purposes (see legal references).*

** *Statistics Canada. 2017. Microdata Linkage Directive. https://www.statcan.gc.ca/en/record/policy4-1.*

*** *Decree N° 2024-12 of 5 January 2024 creating an automated processing system for personal data called the* "répertoire statistique des individus et des logements" *(Statistical Register of Individuals and Dwellings, RÉSIL) (see legal references).*

9 If one of the two files is itself a repository, for example if the aim is to try to update the repository, the identification operation is performed only on the other file.

▶ Data Processing Requires an Appropriate Legal Framework

The record linkage of personal data constitutes data processing in the legal sense and must be treated as such. A controller must be identified and given responsibility for fulfilling the obligations imposed by the General Data Protection Regulation (GDPR) and the Act on Information Technology and Civil Liberties[10]: verification of compliance with the principles of necessity, minimisation and proportionality[11], registration of their administration in the data processing register, conducting an impact assessment if this processing has certain characteristics (for example if it concerns a very large population or uses sensitive variables, etc.). When the record linkage operation concerns statistical data or is carried out for statistical purposes, the data are placed under the protection of the 1951 Act on Legal Obligation, Coordination and Confidentiality in Statistical Matters[12] (*Redor, 2023*).

▶ A Practice that is Often Tricky

When there is no common identifier in the two files, linking must be performed using variables that make it possible to unambiguously identify individuals.

A file contains multiple types of information that play a different role in a data linkage process. This information can be divided into three categories:

- primary identifying information: this relates to identity traits that are uniquely associated with an individual and which are highly stable over time. For a person, this means their forename(s), surname(s) and place and date of birth[13];
- secondary identifying information: this is information that is not uniquely and permanently associated with an individual, but can help to improve the linking process. Possible examples for a person include their municipality and home address;
- other information is generally not used in a linking process, but constitutes variables of interest for the statistical file produced. However, it can be used during the quality assessment, for example by detecting inconsistencies in the linked records.

This poses several difficulties: this information is not always present in the files and may contain errors or omissions. Comparing it is very costly in terms of computer resources, and this cost increases rapidly with the size of the data.

10 Act N° 78-17 of 6 January 1978 on Information Technology, Data Files and Civil Liberties (see the legal references at the end of the paper).
11 According to the GDPR, "Personal data shall be [...] adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')".
12 Act N° 51-711 of 7 June 1951 on Legal Obligation, Coordination and Confidentiality in Statistical Matters (see the legal references at the end of the paper).
13 For an establishment, the company name and address are primary information.

What Criteria Should Be Used When There Is No Common Identifier?

A first prerequisite is of course having the identifying information in both files, and this information must be comparable, i.e. its semantic content and the way it is shown must be identical. For example, the municipality of residence is available in both files and appears in both cases in the form of a label or a code, based on an identical classification.

In addition to this necessary criterion of data comparability, three main conditions are required to successfully perform record linkage:

- the richness of the information;
- the quality of the information;
- an efficient process for processing a large volume of data.

The information available must be sufficiently extensive, differentiating and precise to distinguish individuals from each other. Knowing a month of birth, for example, is much less informative than knowing a forename (one twelfth of the population shares the same value for a month of birth). The more precise the information, the more readily it can be used to distinguish one individual from another. It will therefore be more beneficial to use a complete date of birth rather than a year, a municipality of birth rather than a French department, etc.

It is also possible to attempt to use more information, by using a home address, in addition to identity traits. However, there may still be cases that are not unequivocal (in the case of homonyms, for example), which are all the more frequent when the data are not very precise or contain errors.

Identifying Data Are Never Perfect: How Can They Be Best Used?

The second difficulty relates to data quality. Any missing, incomplete or erroneous information may adversely affect the linkage quality, by causing false links, or conversely, by "missing" many pairs. In a statistical survey, the identity traits of respondents are not among the variables of interest and may consequently suffer from shortcomings in quality. However, this information is essential, given that a linking process is envisaged.

To overcome these difficulties, statisticians have tools available to them to prepare the data used for identification in order to improve the results of their linking operations.

These tools are not magical and cannot conjure up high-quality information if it is missing or incorrect. However, a data standardisation process can be used, in particular to improve comparability:

- the data are standardised in an identical format in both files: converting December to 12, converting the name of a municipality to its municipality code or using similar case (deletion of accented characters, changing to lowercase, etc. (Cotton and Haag, 2023));
- the information is then structured: standardising an address label (road type, road name, number in the road, repetition index and municipality) or identifying a name in a field that also includes a title; however, this second point is less obvious because it requires analysis of the labels.

Sometimes it is a bit trickier, when a single field in an administrative file contains a person's married name and birth name or the name of several holders of a vehicle registration document, for example. This process is effective in so far as the processing carried out is relatively deterministic (such as grouping Bd, boul and Boulevard together under a single name). However, be careful not to go too far: the desire to delete erroneous data can lead to the deletion of information and, ultimately, the degradation of the process (Koumarianos, 2022). This criticism is sometimes made of phonetic algorithms: their purpose is to neutralise orthographic differences, but they can then mistakenly conclude that Lefebvre and Lefèvre, or Schmidt and Schmitt are two identical entries (Randall *et al.*, 2013).

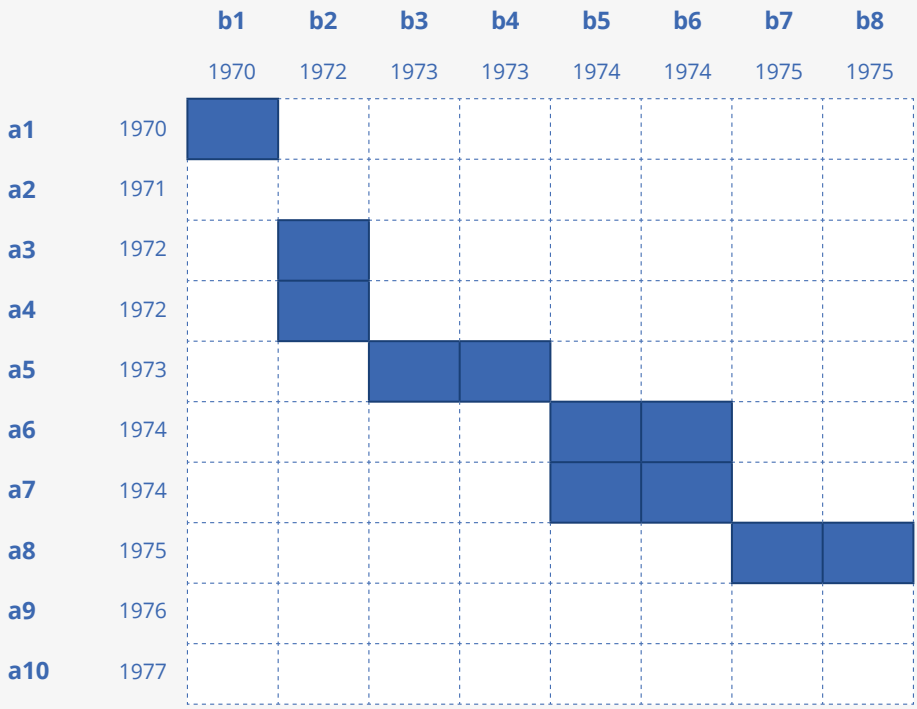
Where errors in the information remain, such as typos or spelling errors, it will often be more effective to deal with these problems later on in the linking process, by using string metrics that take into account these potential typos, rather than making a comparison that relies on strict equality.

▶ Record linkage is a Costly Endeavour; How Can it be Made More Frugal?

> ***This is called blocking: reducing the scale of the problem.***

Finally, a third challenge is performing the process digitally. A linking process amounts to selecting, within two files of sizes N_1 and N_2, the pairs of identical individuals within the set of all possible pairs. This set is very large, with a size equal to $N_1 \times N_2$, whereas the number of pairs of identical individuals is less than or equal to the minimum of (N_1, N_2). Despite technical progress, record linkage remains a large-scale problem. Linking two files of 60,000 lines each means working within a set of 3.6 billion potential pairs. A large number of pairs within the entire set link together two individuals who do not resemble each other at all. It is therefore inefficient to form the theoretical set of all pairs. Statisticians often try to reduce it to a subset of more likely pairs. This is called blocking: reducing the scale of the problem by comparing, for example, only individuals born in the same year (***Figure 2***). The information used for blocking must be of excellent quality. Otherwise, this would result in many missed links.

▶ Figure 2 - Example of Blocking Based on Year of Birth



Reading note: In this fictional example, the aim is to link file A, which contains 10 individuals, with file B, which contains 8.
Blocking is performed based on year of birth, which makes it possible to work on pairs of individuals born in the same year. Instead of forming 80 pairs for comparison purposes (all the squares in the matrix), only 11 pairs are studied (the blue squares).

▶ Successful Record Linkage: a Delicate Balance Between Theory and Technicality, Understanding the Data and Empiricism

A dedicated tool is often used to perform data linking. Statistical tool packages can be used "on a case-by-case basis", but when production is repeated it is common to have dedicated tools for this operation. These tools can be generic and allow any dataset to be linked: in such cases, they generally consist of a set of parameters which must be chosen carefully in order to produce a good result.

Within National Statistics Institutes (NSIs), there are general matching tools that are developed internally or more broadly by another administrative body and which offer a set of tools (comparison and classification functions, a choice of blocking methods)

such as the RELAIS tool developed by ISTAT[14] (Cibella *et al.*, 2012), G-link developed by StatCan[15] (Chevrette, 2011) or SPLink used by the ONS[16] in the UK (Cleaton *et al.*, 2022).

Other tools are more specific and are designed to meet a more targeted need: for example, the tools recently developed within the French Official Statistical Service, such as RAPSODIE[17], which specialises in linking using tax data (Jabot and Treyens, 2018) or InserJeunes (Midy, 2021).

The methods used are not specific, but regular use on a certain type of data leads to selections of rules and parameters that are particularly suited to a specific dataset: correctly coding missing information within a file, for example the SNP (from the French "*Sans Nom Prénom*", meaning "Without Surname Forename") option in an administrative file, or taking into account refuge values for dates of birth on 01/01 or the 15th of each month[18].

▶ Box 2. The Main steps in a Record Linkage Process

Many authors agree on a formalisation of the different steps of a record linkage process (*Christen, 2012*), making a distinction between:

– a data preparation step (which includes quality analysis, standardisation of variables, etc.);

– a pair forming step that takes into account volume issues and optimises the subset created;

– a step in which individuals within pairs are compared, which uses varyingly complex functions to calculate similarity or distance;

– a classification step that selects the retained pairs and discards the rejected pairs;

– an evaluation step that is always necessary and which sometimes leads to the previous steps being modified.

At each step of a record linkage process (***Box 2***), expertise and understanding of the data are vital and often improve the results of the process. There is no turnkey solution that is suitable for all linking operations. It is necessary to take into account the quality of the data and their characteristics when selecting the relevant method and choosing the correct parameters for the comparison and classification functions.

> ***Record linkage is a process that requires granular adjustments of a set of parameters.***

Whatever tool is chosen, record linkage is a process that requires granular adjustments of a set of parameters, which are often approached in an iterative and empirical manner.

14 RELAIS: Record Linkage At ISTAT; ISTAT: Italian national statistics institute.
15 G-Link: Generalized System for Record Linkage; StatCan: Canadian national statistics institute.
16 SPLink: Probabilistic Record Linkage At Scale; ONS: Office for National Statistics, United Kingdom national statistics institute.
17 Reconciliation of social, survey and tax data.
18 These are values in the definition domain of the variable concerned, which are sometimes assigned in cases of non-response. Thus, the date of 01/01 is very often assigned when a date of birth is unknown.

▶ Comparing the Information, then Selecting the Pairs: the Core Aspect of a Linking Process

The rest of the paper does not provide a detailed description of the comparison and classification steps but does give an overview of the key steps of a record linkage process (Christen, 2012, and Malherbe, 2023 for a more comprehensive presentation).

After identifying a set of potential pairs, for example after a blocking step, those pairs are classified. For each of them, the two linked records are compared. This makes it possible to determine whether this is a pair of individuals who are definitely identical, definitely different or whether a doubt remains. There is a wide range of classification methods. They differ in various respects, in particular the greater or lesser degree of automation in the definition of parameters and the use or non-use of a set of annotated pairs[19]. The "probabilistic" approach is characterised by a relatively high degree of automation, together with the use of machine learning[20], whereas other approaches require some key parameters to be defined manually, with the benefit of the statistician's expertise and understanding of the data.

▶ Deterministic Methods: System of Rules...

This method consists in linking the two files over several steps, starting with strict rules and then gradually relaxing the constraints. Individuals linked in one step are no longer considered for the subsequent steps.

> ***Several steps, starting with strict rules and then gradually relaxing the constraints.***

The first step is usually exact matching: if all the linking variables of a line in file A are identical to those of a line in file B (for example, same surname, forename, date and place of birth), the two lines are linked. The following steps allow for slight differences and become increasingly relaxed. Allowing for such differences can be done either by simply excluding a field from the comparison or by imposing a more flexible constraint than an exact match. As the idea is to gradually relax the constraints, it is generally the least selective fields or those containing the most errors that are relaxed first, for example the date of birth rather than the surname: less information is lost and erroneous data that can lead to mismatching is removed.

19 These are pairs for which, most commonly after human observation, the following information is provided and sometimes commented upon: identical individuals, different individuals, impossible to decide.

20 Machine learning is a field of study of artificial intelligence that aims to bestow upon machines the ability to "learn" from data, via mathematical models.
This approach has not proved its worth in the context of linking operations due to the asymmetrical nature of the classification problem on the one hand (statisticians are trying to find n pairs among a dataset of size n^2), and due to the low number of variables on the other (Malherbe, 2023).

▶ ... or Weighted Sum of String Metrics...

Another approach consists in calculating string metrics for each identifier field and then aggregating them to obtain an overall string metric for each pair. A text string metric is a number that represents the distance between two words or texts. Conventional string metrics used for record linkage are based on the Levenshtein distance[21] or the Jaro-Winkler distance[22] (Herzog *et al.*, 2007). The overall similarity of each pair is then obtained using a weighted sum of the similarities of each field. The weights associated with the different variables are defined empirically by the statistician, on the basis of their varyingly selective nature and their quality.

This method is almost always used together with the selection of a threshold. In such cases, a pair is linked only if its overall similarity exceeds the threshold.

▶ ... or Search Engine Linking: an Effective Tool for Managing Large Volumes of Data

A third, much less conventional approach is to use a text search engine, such as Elasticsearch[23] or Solr. This type of tool is primarily designed to efficiently search for information in a very large set of texts, such as all products on an e-commerce site for example. However, it can be very useful for record linkage, especially when the files are very large.

From a practical point of view, record linkage using a search engine is performed quite differently to previous approaches. The first step is to index one of the two files, usually the largest. This operation consists of storing the data in this file such that searches for information contained therein are highly efficient. The data structure used in this framework is called an inverted index[24]. The second step involves performing queries, which means searching in this index for a match for each individual in the other file. The tool gives all the individuals corresponding to the query and classifies them using a relevance score. Search engines are highly flexible in terms of the definition of queries, giving the user considerable freedom to decide which filters and elements are included in the relevance score. This approach can take place in addition to an initial phase of exact matching, which makes it possible to reduce the size of the files to be processed.

21 The Levenshtein distance is a distance, in the mathematical sense of the term, which provides a measure of the difference between two strings. This is equivalent to the minimum number of characters that must be deleted, inserted or replaced to move from one string to another.

22 The Jaro-Winkler distance measures the similarity between two strings. This is a variant that was proposed in 1999 by William E. Winkler, stemming from the Jaro distance which is mainly used for detecting duplicates.

23 For more information on Elasticsearch, see the paper entitled "The Non-Significant Statistical Code (CSNS): a service to facilitate file linking", (Bénichou *et al.*, 2023).

24 For each word found in a text, this structure provides the list of documents that contain it.

This approach has proved its worth as it is used, in particular, in relation to the non-significant statistical code (Bénichou *et al.*, 2023). It will also be used for RÉSIL. It is important to note that it is better suited to identification tasks, i.e. when searching for individuals in a register or in an almost exhaustive file on the population of interest. In such cases, it is more a question of carrying out many "individual" searches, independent of each other, whereas a more conventional linking process will examine two datasets taken in their entirety

▶ The Probabilistic Approach

The probabilistic approach comes from an established mathematical framework (Fellegi and Sunter, 1969). The principle is to assign to each pair a probability of corresponding to a single individual, calculated using a set of parameters. These parameters are not chosen manually, but estimated directly on the basis of the pairs of individuals to be linked (Winkler, 2000). They represent the ability of the various identifying variables to differentiate between pairs. A match on the surname is thus a better clue for linking a pair of individuals than a match on gender. The parameters of the probabilistic linkage model reflect this information using a conditional probability denoted u. This parameter represents the probability of observing the same value in a given field, with the knowledge that the two individuals in the pair are different. For example, the value of this probability for the month of birth would be approximately 1/12.

In addition, the quality of the various identifying variables is taken into account in the parameters using the probability m, which is defined as the probability of observing the same value in a given field within a pair of identical individuals. If the data were of perfect quality, this value would always be 1, but this is rarely the case. The order of 1-m can then be interpreted as the error rate for a given field.

Once the parameters u and m have been estimated for each variable, it is possible to obtain a probability of corresponding to the same individual for each pair. The decision rule then entails comparing this probability with a threshold defined by the statistician to decide whether or not to link the pair. The threshold value should be set on the basis of the objective pursued and the type of error deemed acceptable. If the threshold is high, little risk is taken regarding the pairs that are linked, but there is a risk of missing them. Conversely, if the threshold is low, the match rate is high, but there is a risk of linking pairs incorrectly (see the paragraph on the status of pairs below).

The probabilistic linkage method relies on the data themselves to estimate the u and m parameters. This makes it possible to take into account the informative nature of each variable, without requiring detailed knowledge of the data. However, this method is very costly in terms of compute resources, making it difficult to implement on the volumes of data usually involved in record linkage processes. In terms of quality, the probabilistic method works just as well as deterministic tools adapted to the data (Haag *et al.*, 2022), but cannot compete in terms of computing resources and processing time on large individual data.

Regardless of the method used, record linkage is an imperfect process. Its quality should be assessed both during the development of the process and during its implementation.

▶ Linkage Quality and Statistical Issues

Firstly, assessing quality makes it possible to identify possible areas of improvement for the record linkage. For example, if a specific sub-population is poorly linked, special attention should be paid to it, for example by improving the clean-up and standardisation of the data covering that sub-population. Manual pair examination also makes it possible to identify common errors (such as the reversing of surnames and forenames or the use of old municipality names) and adapt the record linkage to avoid them.

> ***It is important to ensure the quality of the data at the end of the record linkage process and to assess the impact of the process on the study results.***

Many record linkage operations are used for statistical studies. It is important to ensure the quality of the data at the end of the record linkage process and to assess the impact of the process on the study results. This is because insufficient linkage quality leads to inconsistencies in the data on the individuals (false links) or a lack of representativeness (when missing links relate more specifically to certain populations).

The quality of the paired data can be assessed in multiple complementary ways. It is sometimes possible to measure the quality of the process itself, based on the proportion of the population linked or the study of the pairs selected or rejected. It is also desirable to supplement the analysis from a more statistical perspective, by comparing the populations studied before and after record linkage. Is there, for example, the same age structure and the same distribution as throughout the national territory?

▶ Measures of Assessment Based on Pair Status

When record linkage is performed, there are two ways it can go wrong: incorrectly linking two records, that is to say incorrectly believing that two records represent the same person (false positives) or failing to link two records that do represent the same person (false negatives).

When the "true" result is available, the pairs can then be classified into four groups, in accordance with their actual status and the predicted status at the end of the record linkage process (***Figure 3***), such as:

- "Correct decisions":
  - true positives (TPs) are pairs of identical (concordant) individuals who have been linked by the process;
  - true negatives (TNs) are pairs of different (non-concordant) individuals who have not been linked by the process;
- "Incorrect decisions":
  - false positives (FPs) are pairs of different (non-concordant) individuals who have incorrectly been linked by the process;
  - false negatives (FNs) are pairs of identical (concordant) individuals who have incorrectly not been linked (somehow forgotten or not found) by the process.

Although it is a helpful tool, this classification is not a quantitative evaluation of performance. However, it is possible to determine such measurements on the basis of the numbers in each of these categories.

▶ Figure 3 - Pair Status

Predicted status	Actual status:	Identical individuals (concordant pairs)	Different individuals (non-concordant pairs)
		✓	✗
Linked pair	○	True positives ✓	False positives ✗
Unliked pair	▭	False negatives ✓	True negatives ✗

Reading note: 3 pairs are linked, of which 2 are true positives and 1 is a false positive. 22 pairs are rejected, of which 2 are false negatives and 20 true negatives.

▶ Two Indicators Commonly Used to Describe the Quality of a Record Linkage Process

During record linkage, the numbers in each class are extremely unbalanced: for two files of size n, the number of pairs of identical individuals is close to n while the number of pairs of different individuals is approximately n^2. Measurements such as precision and recall, which do not rely on the number of negative pairs, are usually selected (***Figure 4***).

Precision is defined as follows:

$$\text{Precision} \begin{cases} = \dfrac{\text{True positives}}{\text{True positives + False positives}} \\ = \dfrac{\text{Number of linked AND concordant pairs}}{\text{Number of linked pairs}} \\ = \text{Success rate regarding linked pairs} \end{cases}$$

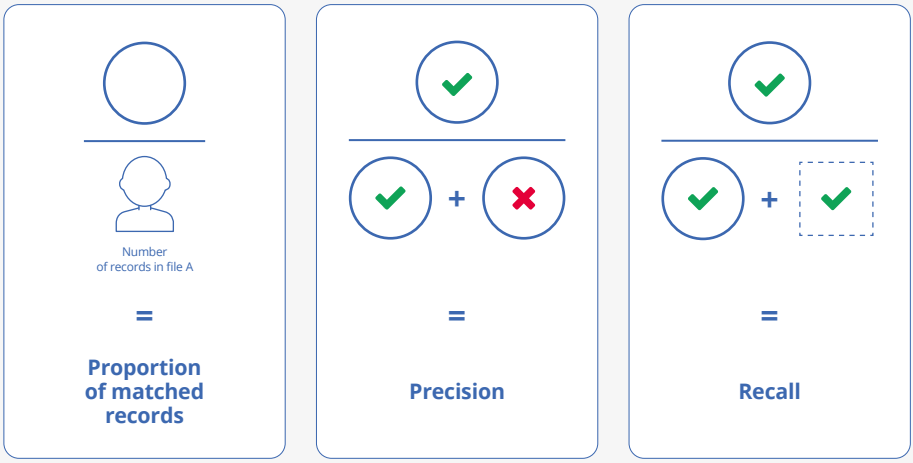
A high level of precision means that errors from this model, when it links a pair, are rare. However, this provides no information on its ability to identify a large number of pairs. As an extreme case, a model linking a single pair correctly would have a perfect precision of 1. However, such a model is not satisfactory. This is why the recall is often used in addition to precision. Recall, also referred to as sensitivity, corresponds to the proportion of positive cases identified as such by the model.

It is defined as follows:

$$\text{Recall} \begin{cases} = \dfrac{\text{True positives}}{\text{True positives + False negatives}} \\ = \dfrac{\text{Number of linked AND concordant pairs}}{\text{Number of concordant pairs}} \\ = \text{Concordant pair identification rate} \end{cases}$$

High recall means that the pairs are easily identified by the model. Here too, in an extreme case, all the potential pairs would be selected, the recall would be 1, but the result would include many pairs incorrectly linked.

▶ Figure 4 - Common Indicators Based on Pair Status



Reading note: In Figure 3, the proportion of linked records is 60% (3/5), the precision is 66% (2/3) and the recall is 50% (2/4).

▶ The End Use of the Linked Data Leads to Trade-Offs Between Precision and Recall

In order to assess the quality of a record linkage process, it is necessary to define objectives and to make a trade-off between false negatives (or missed concordant pairs) and false positives (non-concordant pairs accepted). The concept of quality is always linked to the intended use. Thus, when record linkage techniques are used for operational purposes (for example, in the context of administrative management operations), a great deal of attention is paid to each individual result and, most often, the aim is to avoid false positives (greatest possible precision). In a statistical context, precision is desirable, but it is also desirable to avoid bias in terms of representativeness caused by inadequate recall. It is impossible to be perfect in both respects: if recall and the match rate are increased, then the precision deteriorates, and vice versa. Depending on the intended use of the linked data (data enrichment, coverage assessment, etc.), a decision is made regarding the expected level of precision.

The Tools Needed to Assess the Quality of a Record Linkage Process Based on the Quality of the Pairs

While the proportion of linked records is very easily calculated for any record linkage operation, the same cannot be said of most other measurements of quality. These require additional information about the linked files, usually a sample of annotated pairs. In the best case scenario, we have a gold standard or benchmark file. This is a sample of pairs that is representative of the files to be linked and the actual status of which is known. The way in which such a sample is obtained differs in accordance with each situation.

In most cases, however, there is no gold standard and it is therefore necessary to add a manual annotation step to classify a set of pairs, involving a human observer: is the pair suggested by the process a "valid" pair?, or has the process linked it incorrectly?, or is it impossible to decide? This manual annotation step takes time and can therefore prove to be very costly.

There are other ways to, at least partially, measure these quantities, such as by using a sub-population for which the true match status is, in principle, known or by observing the consistency of the linked data (Doidge et al., 2020). During a record linkage process, it is common to manually annotate a sample of pairs, usually chosen from a subset of pairs with an "uncertain" status; this is the case for pairs for which the similarity is close to the rejection threshold, in order to evaluate the rate of true and false positives on either side of that threshold. If the vast majority of rejected pairs are false negatives, the threshold is modified in order to accept these incorrectly rejected pairs, with the trade-off being a small number of additional false positives.

Assessing the Quality of a Record Linkage Process Based on its Impact on the Data

While the above indicators are used to assess the level of quality of a record linkage process, they are not always easy to measure by end users, especially since the process is sometimes performed by a third-party service. This is often the case, for the purpose of protecting personal data in particular or due to the technical nature of certain operations (data preparation, configuration of the matching tool, etc.).

When the third-party service performs the record linkage, it is aware of all the linking variables and is therefore able to assess the quality of the process using the methods mentioned above. This is not generally the case for end users, who do not have the linking variables (identity traits in particular). Therefore, it is desirable for the third-party service performing the record linkage to produce and send to end users assessments of its process and quality indicators associated with the corresponding result. These measurements are important, but they are not sufficient to assess the statistical impact of record linkage errors.

> ***Assessing the quality of a record linkage process therefore is not the sole responsibility of the entity that carries out the process, but also relies on the complementary work of the one or more services that use the linked data.***

This is because users have a greater number of variables, the variables of interest, which they use to produce statistics (for example, qualifications, profession, income level, etc.). This information makes it possible to assess the impact of the record linkage in a statistical manner, through its impact on the population of interest, in particular through distributions or statistics of the variables of interest. Users can thus check the representativeness of the linked population in relation to the source file: for example, has the age structure of the population been distorted?

This second, more statistical and use-oriented level of analysis is essential to assess any bias introduced by the record linkage process. Therefore, if a statistician is aware of a lack of representativeness in the linked population, they can use adequate statistical processing, as is the case when processing any statistical source.

Assessing the quality of a record linkage process therefore is not the sole responsibility of the entity that carries out the process, but also relies on the complementary work of the one or more services that use the linked data.

▶ Conclusion

Data record linkage has been expanding in official statistics in recent years, driven both by a growing demand for enriched data and by the increasing availability of administrative data and computational resources.

It is essential to various statistical processes and will be at the heart of the RÉSIL[25] programme.

It is important to assess the quality of this record linkage, not only during the execution of the process but also during subsequent uses of the linked data.

While there are identified tools and methods for performing record linkage, they must be supplemented by data analysis and statisticians' expertise to select the most appropriate parameters for the datasets concerned.

25 See the paper by Olivier Lefebvre on RÉSIL in this issue.

▶ Legal References

- Act N° 51-711 of 7 June 1951 on Legal Obligation, Coordination and Confidentiality in Statistical Matters. At: *Légifrance website*. [online]. Updated on 25 March 2019. [Accessed on 22 February 2024]. Available at: *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

- Act N° 78-17 of 6 January 1978 on Information Technology, Data Files and Civil Liberties. At: *Légifrance website*. [online]. Updated on 21 February 2024. [Accessed on 22 February 2024]. Available at: *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460*.

- Decree No 2016-1930 of 28 December 2016 simplifying the advance formalities relating to processing for statistical or research purposes. At: *Légifrance website*. [online]. Initial version. [Accessed on 22 February 2024]. Available at: *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033735139/*.

- Decree No 2024-12 of 5 January 2024 creating an automated processing system for personal data called the "*répertoire statistique des individus et des logements*" (Statistical Register of Individuals and Dwellings, RÉSIL). At: *Légifrance website*. [online]. Initial version. [Accessed on 22 February 2024]. Available at: *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000048866207*.

▶ Bibliography

- ANDRÉ, Mathias and MESLIN, Olivier, 2022. Property wealth of households: Lessons from the use of comprehensive administrative data sources combination. In: *Courrier des statistiques*. [online]. 20 January 2022. Insee. N° N7, pp. 107-125. [Accessed 20 february 2024]. Available at: *https://www.insee.fr/en/information/7722560?sommaire=7722566*.

- BÉNICHOU, Yves-Laurent, ESPINASSE, Lionel and GILLES, Séverine, 2023. The Non-Significant Statistical Code (CSNS): a service to facilitate file linking. In: *Courrier des statistiques*. [online]. 30 June 2023. Insee. N° N9, pp 64-85. [Accessed 20 February 2024]. Available at: *https://www.insee.fr/en/information/8232634?sommaire=8232646*.

- CHEVRETTE, Antoine, 2011. G-Link: A Probabilistic Record Linkage System. In: *NORC Conference Proceedings*. [online]. May 2011. [Accessed 20 February 2024]. Available at: *https://www.norc.org/content/dam/norc-org/pdfs/G-Link_Probabilistic%20Record%20Linkage%20paper_PVERConf_May2011.pdf*.

- CHRISTEN, Peter, 2012. *Data Matching–Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. 5 July 2012. Springer. ISBN 978-3642311635.

- CIBELLA, Nicoletta, SCANNAPIECO, Monica, TOSCO, Laura, TUOTO, Tiziana and VALENTINO, Luca, 2012. Record Linkage with RELAIS: Experiences and Challenges. In: *Site de Istat*. [online]. [Accessed 20 February 2024]. Available at: *https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/relais*.

- CLEATON, Mary, HALL, Johanna, SHIPSEY, Rachel, WHITE, Zoe and XHAFERAJ, Kristina, 2022. A case study of using Splink: Census duplicate matching. Proceedings of Statistics Canada Symposium 2022. In: *Plateforme open source GitHub de l'Office for National Statistics*. [online]. [Accessed 20 February 2024]. Available at: *https://github.com/Data-Linkage/Splink-census-linkage/blob/main/SplinkCaseStudy.pdf*.

- COTTON, Franck and HAAG, Olivier, 2023. Integrating administrative data into a statistical process - Industrialising a key phase. In: *Courrier des statistiques*. [en ligne]. 30 June 2023. Insee. N° N9, pp 104-125. [Accessed 20 February 2024]. Available at: *https://www.insee.fr/en/information/8232639?sommaire=8232646*.

- DEMOTES-MAINARD, Magali, 2019. Élire, an Ambitious Project to the Benefit of the Single Electoral Register. In: *Courrier des statistiques*. [online]. 27 June 2019. Insee. N° N2, pp. 58-71. [Accessed 20 February 2024]. Available at: *https://www.insee.fr/en/information/4195090?sommaire=4195125*.

- DOIDGE, James, CHRISTEN, Peter and HARRON, Katie, 2020. Quality assessment in data linkage. National Statistician's Quality Review. In: *Site de UK government*. Updated 16 July 2021. [online]. [Accessed 20 February 2024]. Available at: *https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage*.

- ESPINASSE, Lionel and ROUX, Valérie, 2022. The National Register for the Identification of Individuals (RNIPP) at the core of French administrative life. In: *Courrier des statistiques*. [online]. 29 November 2022. Insee. N° N8, pp. 72-92. [Accessed 20 February 2024]. Available at: *https://www.insee.fr/en/information/6692672?sommaire=6692684*.

- FELLEGI, Ivan P. and SUNTER, Alan B., 1969. A theory for record linkage. In: *Journal of the American Statistical Association*. Vol. 64, N° 328, pp. 1183-1210. [online]. [Accessed 20 February 2024]. Available at: *https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf*.

- HAAG, Olivier, KOUMARIANOS, Heidi and MALHERBE, Lucas, 2022. Probabilistes ou déterministes, des méthodes d'appariements au banc d'essai du programme Résil. In: *Site des JMS de l'Insee*. [online]. JMS 2022. [Accessed 20 February 2024]. Available at: *https://journees-methodologie-statistique.insee.net/probabilistes-ou-deterministes-des-methodes-dappariements-au-banc-dessai-du-programme-resil/*.

- HACHID, Ali and LECLAIR, Marie, 2022. Sirus, the business register for statisticians. In: *Courrier des statistiques*. [online]. 29 November 2022. Insee. N° N8, pp. 115-130. [Accessed 20 February 2024]. Available at: *https://www.insee.fr/en/information/6692680?sommaire=6692684*.

- HERZOG, Thomas N., SCHEUREN, Fritz J. and WINKLER, William E., 2007. Data Quality and Record Linkage. In: *Researchgate*. [online]. January 2007. [Accessed 20 February 2024]. Available at: *https://www.researchgate.net/publication/220695391_Data_Quality_and_Record_Linkage*.

- JABOT, Patrick and TREYENS, Pierre-Eric, 2018. Proposition d'un nouveau processus d'appariement au Pôle Revenus Fiscaux et Sociaux (RFS). Une application à l'enquête CARE. In: *Actes des journées de méthodologie statistique 2018*. [online]. [Accessed 20 February 2024]. Available at: *https://journees-methodologie-statistique.insee.net/lappariement-denquetes-avec-des-donnees-administratives-sociales-ou-fiscales/*.

- KOUMARIANOS, Heidi, 2022. Impact du nettoyage des données sur la qualité d'un appariement. In: *Site des JMS de l'Insee*. [online]. JMS 2022. [Accessed 20 February 2024]. Available at: *https://journees-methodologie-statistique.insee.net/impact-du-nettoyage-des-donnees-sur-la-qualite-dun-appariement/*.

- MALHERBE, Lucas, 2023. Appariements de données individuelles : concepts, méthodes, conseils. In: *Documents de travail n° M2023/03*. [online]. 3 July 2023. [Accessed 20 February 2024]. Available at: *https://www.insee.fr/fr/statistiques/7644535*.

- MIDY, Loïc, 2021. A matching tool using indirect identifiers - The example of the information system on the integration of young people (système d'information sur l'insertion des jeunes). In: *Courrier des statistiques*. [online]. 8 July 2021. Insee. N° N6, pp. 82-99. [Accessed 20 February 2024]. Available at: *https://www.insee.fr/en/information/7673384?sommaire=7673393*.

- RANDALL, Sean M., FERRANTE, Anna M., BOYD, James H. and SEMMENS, James B., 2013. The effect of data cleaning on record linkage quality. In: *BMC Medical Informatics and Decision Making*. Vol. 13, nº 1, pp. 64. [online]. 5 June 2013. [Accessed 20 February 2024]. Available at: *DOI 10.1186/1472-6947-13-64*.

- REDOR, Patrick, 2023. Confidentialité des données statistiques : un enjeu majeur pour le service statistique public. In: *Courrier des statistiques*. [online]. 30 June 2023. Insee. Nº N9, pp. 46-63. [Accessed 20 February 2024]. Available at: *https://www.insee.fr/fr/information/7635823?sommaire=7635842*.

- ROSENBAUM, Paul R. and RUBIN, Donald B., 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. In: *Biometrika*. Vol. 70, Nº. 1, pp. 41-55. [online]. April 1983. [Accessed 20 February 2024]. Available at: *https://www.jstor.org/stable/2335942*.

- Statistique Canada. 2017. Directive on Microdata Linkage. In: *site de Statistique Canada*. [online]. [Accessed 20 February 2024]. Available at: *https://www.statcan.gc.ca/en/record/policy4-1*.

- WINKLER, William E., 2000. Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. In: *Statistical Research Report Series*. RR2000/05, US Bureau of the Census. [online]. 4 October 2000. [Accessed 20 February 2024]. Available at: *https://courses.cs.washington.edu/courses/cse590q/04au/papers/WinklerEM.pdf*.