

Statistics based on administrative data: towards a general framework

Working Paper

N° 2025-03 - September 2026



Heidi KOUMARIANOS
Pascal RIVIÈRE

M 2025/03

**Statistics based on administrative data:
towards a general framework**

**Heidi KOUMARIANOS
Pascal RIVIÈRE**

INSEE

September 2026

Note : This working paper was originally published in French in June 2025.
The English version is published in September 2026

Acknowledgements:

The authors would like to thank Hélène Chaput, Olivier Haag, Pierre Lamarche, Joël Bizingre, Mylène Chaleix, Elise Coudin, Pascale Breuil, Eric Lesage and Corinne Prost for their thorough reading of, and comments on, this report.



Direction de la méthodologie et de la coordination statistique et internationale
Département des Méthodes Statistiques - Timbre L001 -
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : <http://www.insee.fr>
Ces documents de travail ne reflètent pas la position de l'Insee et n'engagent que leurs auteurs.
Working papers do not reflect the position of INSEE but only their authors' views.

Abstract

The use of administrative data to produce statistics, and the questions arising from this approach, are nothing new to official statisticians. Nonetheless, they have not yet given rise to a methodological framework comparable to that available for survey data. This use of such data is sometimes assumed, incorrectly, to be straightforward, since the data already exist and statisticians do not need to conduct a separate data collection process to obtain them.

However, this apparent simplicity is misleading: the fact that there is no data collection process also means that we lose control over the methods used to gather the data, their time frame, the definition of the variables, the classification systems used, and of course the opportunity to verify the data... Ultimately, these new data sources require us to rethink a whole swathe of the statistical production process, in order to address the new questions they raise.

By their very nature, administrative data are associated with a specific environment, with its own objectives, language, categories and dynamics. As by-products of administrative activities, they are not, strictly speaking, self-contained data suitable for statistical use. Indeed, they may turn out to be far removed from the environment in which data users usually operate. Using administrative data for statistical purposes raises specific quality issues, since quality as a concept is always dependent on intended use. The challenge is to successfully make the transition from the administrative realm to the statistical realm, detaching the data from one context and re-attaching it to another. Such an undertaking necessarily involves certain difficulties and tensions, encompassed in the notion of “data friction”.

In order to make this transition meticulously and effectively, we need an analytical framework. That framework must address five dimensions: objects (or statistical units), population and scope, variables, domains (or classification) and time frames.

In this paper we propose an approach based upon just such a framework, divided into three phases: acquisition, transformation and statistical processing.

- The acquisition phase takes as its starting point the data extracted from an administrative information system, an IS not designed for statistical purposes because it is heterogeneous, incomplete and constantly changing, and designed for specific operational or administrative purposes. The goal is to gather, document, filter and freeze the data, establishing a usable “administrative source”.
- The aim of the transformation phase is to make the transition from the administrative realm to the statistical realm.
- The statistical processing phase is more familiar, closely resembling the process applied to survey data.

For each of these phases, we highlight the importance of “feedback loops” based on the principle of data tracking: in order to guarantee the quality of the data, and thus of the resulting statistics, we sometimes need to trace problems back through the process. This is all the more difficult since, unlike when working with survey data, we do not have total control over the concepts. Numerous checks are thus required, at multiple levels: if an anomaly is detected, we follow its trail back into the earlier phases of the process, back to the source if necessary. The nature of these loops is very different from one phase to the next. Ultimately, because of the quality and explainability we expect from our results, these phases do not follow one another in a purely linear manner.

This document proposes a general framework, but its practical implementation may require several adjustments to overcome practical constraints (lack of volume, or lack of time, for example). The framework is designed for use with a single administrative source, but it can be expanded: the principles established herein are largely applicable to private-sector data too, notwithstanding certain additional difficulties (scope, confidentiality, agreements, cost etc.). These principles can also be applied in a multi-source context, combining administrative and survey data, a configuration which raises further questions of its own.

Keywords: administrative data, quality assessment, metadata, process, transformation, feedback loop, GSBPM

JEL Classification: C80, C40

Note to readers of the English version :

This document was first written in French.

Throughout the document, we provide as much context as possible on France and on the statistical processes used by our national statistical institute.

A list of acronyms can be found at the end of the document in Annex 4.

Table of contents

Introduction.....	3
A. Defining characteristics of survey data and administrative data.....	5
1. Using surveys to produce statistics: implicit and explicit properties.....	6
1.a. The traditional mathematical representation... and analytical framework.....	6
1.b. A two-stage observation framework.....	8
1.c. The importance of multi-level verification.....	9
1.d. Key characteristics of statistics derived from surveys.....	10
1.e. Representing the process: the GSBPM.....	12
2. Using administrative data to produce statistics.....	13
2.a. Some characteristics of administrative data.....	13
2.b. The concept of data... and a new analytical framework.....	15
2.c. Administrative data for statistical purposes.....	16
2.d. Statistics derived from administrative data vs statistics derived from surveys.....	20
2.e. Consequences for the process, compared with surveys.....	21
3. Administrative information systems and administrative sources.....	24
3.a. The concept of the administrative information system.....	24
3.b. Freezing data: a necessary construct implying some decisive choices.....	25
3.c. The “administrative source”: a fragile concept.....	29
3.d. The delicate transition from one domain to another	31
3.e. Statistics based on administrative data: a three-phase process.....	34
B. Our proposed approach.....	36
4. The acquisition phase: evaluating the quality of administrative sources.....	37
4.a. Principles.....	37
4.b. The preparation phase: source, ecosystem, contracts.....	38
4.c. The quality of administrative data: an unstable notion.....	42
4.d. The data receipt and quality assessment phase.....	45
4.e. Feedback loops with administrative entities.....	50
4.f. Repeat and multiple uses.....	52
4.g. Outcomes of the acquisition phase, and moving to the next phase.....	54
5. The transformation phase: making the transition from the administrative realm to the statistical realm.....	57
5.a. Principles.....	57
5.b. Principal operations in the transformation phase.....	58
5.c. A technical process, which we must be capable of replicating.....	63
5.d. Evaluating the quality of transformed data.....	64
5.e. A feedback loop operating on two levels.....	66

5.f. Outcomes of the transformation phase.....	68
6. The statistical processing phase.....	71
6.a. Context.....	71
6.b. The processing operations.....	72
6.c. Quality assessment and feedback: the importance of <i>output editing</i>	75
6.d. Outcomes of the statistical processing phase.....	76
C. Conclusion.....	78
7. Summary and outlook.....	79
7.a. Summary of our proposed approach.....	79
7.b. Implementation.....	83
7.c. Further developments.....	84
Bibliography.....	86
Annex 1: Possible classification systems for administrative data.....	96
Annex 2: Dimensions used by Istat to evaluate the quality of administrative data.....	101
Annex 3: List of administrative data sources cited in this article.....	102
Annex 4: List of acronyms – only in english version.....	104

Introduction

In the 1930s, the development of probability sampling (Neyman 1934) finally made it possible to conduct surveys at a reasonable cost, requiring only a small sample (Ardilly 2006). This approach has since spread far and wide, becoming standard practice for official statisticians, coexisting with censuses (Desrosières 1993). All over the world, a vast wealth of know-how has gradually been accumulated, enriched and consolidated, built upon a corpus of methods and practices (for sampling, questionnaire design, how to handle missing values etc.), recognised by the academic community and officially adopted by international statistical organisations. Little by little, “statistician” came to be recognised as a genuine occupation (Volle 1980).

Meanwhile, national governments all over the world have sought ways to make use of their own administrative data for official statistical purposes, complementing their survey efforts. Some, particularly in Nordic countries, have a long and successful record of integrating administrative data on a massive scale (Wallgren & Wallgren 2007). In France, INSEE has been using administrative data since its inception. Declarations containing social data represent a particularly important source and have been used widely and frequently since the 1950s. On a general level, at a time when it is becoming increasingly difficult to conduct surveys (Meyer, Mok, Sullivan 2015), administrative data have an essential role to play in official statistics. Among other things, they allow us to construct reference registers (Wallgren & Wallgren 2016, Lefebvre 2024), match data with survey responses and thus enrich the array of available data (Koumarianos, Lefebvre, Malherbe 2024, Brion, 2011), position our work within a general system of data acquisition for statistical purposes (Bakker, Van Rooijen, Van Toor 2014), mitigate the problem of nonresponse (Bycroft, Mateson-Dunning 2020), assess sources of bias (D’Aurizio, Papadia 2019), calculate totals in order to calibrate margins (Deville, Särndal 1992) and improve population measurements at a fine level of geographical detail (Zhang 2021).

Administrative datasets may also be used simply to produce statistics, and that is precisely the subject of the present working paper. Our ambition is to clearly define the constituent components of this process, taking into account the specific properties of administrative data. We therefore do not address matters of data linkage, source combinations, precision estimation and the like. We also omit a whole series of topics frequently associated with administrative data in the academic literature: confidentiality, social acceptability, ethics, legal framework etc.

When using administrative data to produce statistics, we find ourselves in an unusual situation where “the data already exist”. Initially, we might assume that this would make our work considerably easier than conducting surveys: no questionnaire to design, no sample to select, no collection process to implement. In reality, however, it is a fundamentally different context, a context that gives rise to very different requirements and in which generic processes are not particularly clear-cut, a far cry from the world of surveys. For surveys, the General Statistical Business Process Model (GSBPM) now serves as a standard model which is recognised and shared by national statistical institutes (UNECE 2019); but the GSBPM is only partially relevant to cases involving administrative sources.

We have also noticed that, although there are many papers dealing with statistics produced using administrative data from specific datasets,¹ and a handful of very general papers (Hand 2018, Zhang 2012, Rouppert 2005), very few general conceptual frameworks exist. However, using new sources instead of, or alongside, survey data is a change far more profound than it may seem (Elbaum 2018), a change capable of delivering real benefits, but one which also carries certain limitations, and requires us to challenge our own assumptions and rethink some of our working methods. Salgado & Oancea (2020), for example, offer a detailed comparison of the different types of sources

¹ This has been a frequent topic of discussion at INSEE’s ‘Statistical Methodology Days’, for example.

at our disposal (survey data, administrative data, digital data), considering their methodological implications while also looking at other dimensions such as quality, technological implications and accessibility.

In this working document, we sketch the possible contours of such a framework. We begin with three chapters on the fundamentals of this subject, defining survey data and administrative data, followed by three more chapters dealing with our proposed approach, before concluding with a summary.

Chapter 1 deals with “traditional” statistics derived from surveys, as we seek to identify their constituent characteristics, both explicit and implicit.

Having established this analytical framework, in Chapter 2 we proceed to look at “administrative data” and identify their principal characteristics: we can thus compare and contrast these properties with those identified for survey data, in order to analyse their respective benefits and limitations. Of course, these data never exist in isolation: they belong to what are generally known as “administrative sources”, although the precise definition of this term is not always clear.

In Chapter 3 we demonstrate that administrative sources are in fact produced by “administrative information systems”, incorporating their own domain-specific perspectives which are often overlooked and difficult to formalise, and yet crucially important. Passing from the administrative realm to another sphere altogether (statistics or research, for example) is not possible without a certain degree of friction and several transformations, all of which raises questions about the usefulness of the label ‘raw data’. From these reflections, we arrive at a representation of the process of producing statistics using administrative data which we divide into three main phases: acquisition, transformation and statistical processing.

Our fourth chapter is devoted to the acquisition phase, that part of the process wherein we retrieve administrative data sources that conform to the agreed specifications, are correctly structured and are sufficiently documented. We insist particularly upon the notion of quality, of quality assessment and improvement, and the importance of establishing a feedback loop to ensure continued improvement. Acquisition is probably the phase which is least understood and least formally defined in the existing literature.

Chapter 5 concerns itself with the transformation phase, which enables us to make the switch from the administrative realm (which has its own distinctive language, operational objectives and semantics) to the statistical realm. We present this phase as a series of steps which we seek to automate, although this does nothing to dent the importance of verification of a different nature. By the end of this phase, we should find ourselves with a dataset fit for its intended statistical use.

The statistical processing phase is described in a far more concise manner in our sixth chapter. Broadly speaking, we use the familiar statistical processes one would expect when working with survey data, simply taking certain precautions to identify areas where the administrative nature of the data might have certain consequences.

In the final section of the paper, having summarised our approach once more, we consider its practical implications and reflect upon potential avenues for further development.

A. Defining characteristics of survey data and administrative data

1. Using surveys to produce statistics: implicit and explicit properties

For Lessler & Kalsbeek (1992), a survey is a “scientific study of an existing population of units typified by persons, institutions, or physical objects.”

INSEE, and national statistical institutes in general, have extensive experience of both surveys and the production of statistics based upon survey data. Over the years, methods have been designed, documented and enriched, and on these foundations tools have been developed and organisational structures put in place, particularly when it comes to collecting data from households or businesses by means of surveys. A substantial body of practical knowledge and expertise has been accumulated.

In order to comprehend and demonstrate these techniques clearly, we deliberately limit our analysis to a very simplified representation which allows us to explain and compare statistics derived from surveys and statistics derived from administrative data.

1.a. The traditional mathematical representation... and analytical framework

The starting point is simple: we find ourselves in a situation where we want to estimate the total value of a variable for a given target population. For example, the total turnover from a sector of the economy, or the number of unemployed people in a given region. We often also wish to calculate a mean value, which involves dividing this total by the number of individuals concerned (e.g. mean age of residents in a given settlement), which is equivalent to a ratio of aggregates.

To keep things simple, let's start with a total. Say that we wish to estimate the total value of variable X for population U in reference period or date t :

$$X^t = \sum_{i \in U} X_i^t \quad (1)$$

When conducting a survey, we often know the size of the population U , for which we have a sampling frame. We randomly select a sample S from within that frame, and since we know the probability of inclusion, survey theory allows us to construct an estimator which brings us close to the total value we are seeking to establish:

$$\hat{X}^t = \sum_{i \in S \subset U} w_i X_i^t \quad (2)$$

where, as in the Horvitz-Thompson estimator (1952), the survey weights w_i are inversely proportional to the probabilities of inclusion.

In reality, equation (1) is too general. Firstly, statistics are calculated for a specific “cell” in the spreadsheet, for example all people between the ages of 25 and 34, or for all businesses in the retail sector: which is to say that they generally focus on a specific sub-population (or category) C_k . Furthermore, we are looking to ascertain the total value of X at a given date, e.g. the total turnover of the retail sector in 2023, X^{2023} .

The total we are looking for must therefore look something like this:

$$X_{C_k}^t = \sum_{\substack{i \in U \\ i \in C_k}} X_i^t \quad (3)$$

where t is the reference period/date, for example a year.

In order to estimate this total, we must take the sampling parameters into consideration (including survey weights), as well as the fact that certain values X_i^t are missing or anomalous, requiring us to impute them with the help of an imputation model.² We use X_i^t to denote the imputed value.

The estimator can be written:

$$\hat{X}_{C_k}^t = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i \tilde{X}_i^t \quad (4)$$

In this very simple formula, $\tilde{X}_i^t = X_i^t$ if the value is preserved,³ or else is derived from a model. Of course we could make this more complex, for example by marginal calibration.

A probabilistic representation of survey data makes it technically possible to calculate the variance of the estimator, and also to calculate the total survey error (Groves, Lyberg 2010; Lyberg, Weisberg 2016; Biemer et al. 2017).⁴ Returning to the last equation, sample S and the imputed value X_i^t are random variables.⁵ We can also attempt to estimate changes over time, and here again we can assess the precision of the estimate, bringing in the joint probabilities of inclusion: we know how to quantify the uncertainty, the margin of error.⁶

In order to give this some concrete shape, and actually obtain values for population, observations, category divisions etc., we must put in place an observation process in order to get as close as we possibly can to the different parameters contained in our initial equation. This formula (number 4 above) represents the aggregate we seek to estimate, and shows that we need five components (Figure 1):

1. The unit (i)
2. The target population, or scope (U)
3. The variable (X)
4. The classification (C)
5. The reference period or date (t)

Figure 1 – Framework for the analysis of survey data

² Of course we could choose reweighting over imputation.

³ Which is not the case when imputation is involved.

⁴ The idea of total survey error is to measure the margin of error in a manner which goes beyond simply calculating sampling variance, including other error components (such as imputation error). This subject is worthy of a chapter in its own right, which would overstep the boundaries of this paper. We thus limit ourselves to citing some germane references.

⁵ This is a simplified formula containing the major error components, chief among which is sampling error, but many other forms of error are identified in the academic literature on this topic.

⁶ Hand (2018): “statistics is the technology of extracting meaning from data and of handling uncertainty.”

1.b. A two-stage observation framework

The methodological framework for survey-based statistics is not merely a proliferation of mathematical formulae, however. Broadly speaking, its purpose is to serve as an *observation instrument* of the kind one would expect to find in other fields: archaeology, biology, astronomy etc. (Borgman 2015). Each of these systems has its own specific strengths and weaknesses, as demonstrated by Borgman et al. (2016) in the case of astronomy.

What about using surveys as a means of observing socio-economic realities? The distinguishing characteristic of a statistical observation framework is that it seeks to observe “macro” phenomena for an entire population (employment, poverty, crime, health etc.). For this reason, it operates in two phases: firstly, gathering and defining a “large” number of individual observations, in order to subsequently put them to use for statistical ends, not least by establishing aggregate data which are, in some respects, macroscopic observations.

In order to obtain individual observations, we make use of an *instrument* of observation, which in our case is called a *questionnaire* (in biology, for example, it would be a microscope). The *system* of observation, as we shall call it, is what produces these individual observations, for the units contained in the *sample*.

For any survey, this system requires a long phase of preliminary design work, followed by a construction phase (as if astronomers had to rebuild their telescopes for each new observation), and finally the operational implementation phase, which we call *data collection*, during which data are collected from the target population.⁷

We thus have three consecutive phases:

- Design: the questionnaire (questions permitting us to ascertain X_i), the sampling frame, sampling (the units denoted by i), choice of data collection method etc.
- Construction: developing the data-collection instruments, fine-tuning the protocol⁸
- Implementation: actually collecting the data.

The first level of observation corresponds to the entire mass of individual data gathered: this is the output of the data collection phase, contained in one or more files. It is, in some respects, an intermediate product defined by its tabular format: a “rectangle” with p variables for n individuals. Several iterations are required to achieve this. In practice, this is a slightly more complex operation, but for our purposes this simplified version is more than sufficient.

The second level of observation takes this file, or rectangle, as its starting point. It yields results in the form of a set of statistics (aggregates, means, ratios, quantiles, distribution properties etc.).⁹ The first task is to complete and (almost) finalise these results: imputing missing or anomalous data points, as well as performing classification (C_k),¹⁰ an operation whose importance was demonstrated above; for example, coding for cause of death (Coudin, Robert 2024), using multiple labels and other information, requires a considerable amount of interpretation. It is on these foundations that the statistics themselves are built, using statistical methods of estimation, processing nonresponses, marginal calibration etc.

⁷ Strictly speaking, we must make the distinction between collection units (units *from* whom we collect data) and statistical units (*about* whom we collect data): in business surveys, the collection units may be accountancy firms.

⁸ Including complex protocols, such as multimodal approaches (Beck et al 2022).

⁹ Of course there are other possible uses, particularly econometric uses, but for simplicity’s sake we will focus exclusively on descriptive statistics.

¹⁰ Which may be present in the sampling frame (for example the statistical classification of economic activities for business surveys), or obtained from first-level observations.

The fact that this processing is based upon a statistical model enables us, at least in principle, to estimate variance and its different components. In other words, equipped with all of the levers of the statistical production process, and working within a mathematical framework, we are capable of measuring the various dimensions of error. Survey statistics allow us, and therein lies their great strength, to keep error under control (Deville 1997). It should be noted, however, that for statistics which continue to evolve, all of these mechanisms demand a certain amount of consistency from the analytical framework over time: variables X , units i , categories C_k , population U etc.

1.c. The importance of multi-level verification

One central pillar of the survey process is the vast amount of verification work involved, both automatic and manual, at different points in the process. Why is verification so important? Because official statistics tend to attract many users, who are – entirely naturally – likely to compare them with other statistics (comparing them with previous periods in time, or with other, equivalent statistics). Allowing for such geographical and temporal comparisons,¹¹ and thus demonstrating consistency, is one of the core demands of official statistics, along with the capacity to explain and justify statistics in the interests of transparency.

It is thus entirely understandable that there are so many forms of checks of different types and conducted at different stages, which may or may not be associated with automatic imputation operations: collectively, they are known as *data editing* (De Waal, Pannekoek, Scholtus 2011).¹² The first that springs to mind is verification at source, performed by the person conducting the survey during the collection process: conducting surveys in person makes it possible to check that the question has been understood correctly, and that respondents are truly sure of their answers, ensuring that responses are correctly interpreted during classification. Electronic collection methods also allow us to perform certain basic checks automatically (checking datatypes, domain belonging, and even consistency between variables), within the collection tool, preventing some errors.¹³

In business surveys, verification by interaction with respondents can be done later on, but always relatively quickly, by calling surveyed companies to check any responses which seem unusual. The work done by survey clerks thus allows us to fine-tune the data, which is essential.

Further checks may also be conducted much later on in the process, during the statistical processing phase, by selectively analysing those units which have a strong impact on the statistics, thus conducting another round of individual verifications (Lawrence, McKenzie 2000).

During this *data editing* process, verification generally begins with automatic checks which allow us to identify anomalies (figures given in Euros instead of thousands of Euros, for example). The process may continue with manual checks if an anomaly is detected whose scale is deemed to be statistically significant, and for which automatic correction could be too large a source of potential error, requiring feedback to correct the anomalous value... or not (results which appear to be anomalous are not always errors). The verification process thus incorporates a *feedback loop* which may result in some data being modified. The principle behind the feedback loop is broadly applicable to all processes and their verification (Wiener 1948). It is one of the core tenets of

11 Comparability is one dimension of the European statistical quality framework.

12 *Data editing* is first and foremost a matter of checking data, automatically or manually, and dealing with the outcomes of this verification – i.e. the fact that data are either confirmed or called into question. In the latter case, manual modifications by the clerks may be necessary, or else automatic processing methods may be used (imputation using models). As such, *data editing* only involves imputation when initial verifications prove to be insufficient.

13 These consistency checks are rarely blocking, but electronic data collection tools impose certain constraints on the format of survey responses.

cybernetics, the science of controlled systems, and is a natural fit for the process of producing statistics.

As such, whether performed during data collection or later on,¹⁴ feedback operates *at the level of individual statistical units*: modifications affect the *microdata*. As we shall see, things are very different when working with administrative data.

At the very end of the survey process, checks are conducted upon the aggregates themselves, a task known as *output editing*. The fact that a given statistic varies considerably over time does not necessarily indicate that an error has been made, but this variation does need to be detected, properly understood and documented, or even broken down into components (variation of stable units, variation linked to demographic changes in the units which make up the scope, etc.). Output editing is sometimes performed informally by survey clerks themselves, as they are often the first users of the results, and well-placed to detect inconsistencies. But even in such cases, the presence of a feedback loop may lead to changes in the microdata.¹⁵

These multiple verifications, which must be used sparingly but sufficiently (Granquist 1997), play an indispensable role in the survey process, with a view to improving the quality of individual data and thus, by extension, of the resulting aggregates. For automatic corrections (imputation in particular), it is also possible to measure their impact on the accuracy of the aggregates.

1.d. Key characteristics of statistics derived from surveys

The mathematical representation given in Section 1a illustrates the various dimensions of surveys (statistical units, variables, scope, categories, time frames), to which we shall return now. Ultimately, the truly defining characteristic of the process of producing statistics from surveys is the necessity of achieving a certain degree of control over the 5 components (Nordbotten 2010)¹⁶ which make up our analytical framework:

- the statistical units to be observed, defined very early on in the process, and which are often (but not always) individuals, households, businesses or establishments,¹⁷
- the scope of observation: defined to ensure that it is meaningful and, preferably, accessible before collection begins.
- For a business survey, for example, we can define the scope with reference to the sector of activity or the size class of the enterprise concerned, but not with reference to the total value of investments, in light of the information available in the sampling frame.¹⁸
- the variables to be observed: defined after consultations with the various stakeholders (data users and providers),¹⁹ via the dedicated channels (Anxionnaz, Maurel 2020),
- the classification principles: shared classification schemes (Guibert, Laganier, Volle 1971), e.g. the occupational classification (Amossé 2020), or classification of legal offences (Camus 2022),

14 A useful modelling option in the GSBPM (cf. § 1.e).

15 Of course it would be wrong to imagine that surveys always take place in such idyllic circumstances. There are many situations in which it becomes impossible to modify the microdata: for example when working with survey data from long-running series. Some tasks (back projection of conceptual changes, new classification schemes etc.) are performed at an aggregated level.

16 Three of these components – unit, population and time – form a classic triptych long regarded as pillars of statistics: Nordbotten mentions them in his 2010 article, which in fact harks back to a paper he wrote in the 1960s.

17 In systems such as the user surveys by the DREES, we address hospital stays with a threefold approach (establishment, physician, patient) – DREES is the french national statistical office of Health minister.

18 The sampling frame is a constraint which could potentially limit the scope of observation.

19 One example would be businesses, represented by federations or employers' organisations.

- finally, two temporal dimensions applied uniformly to all units: the time frame of data acquisition (the period when the data is collected) and the reference time frame for these data (i.e. the date or period to which they pertain).²⁰ This final property is absolutely essential, and constitutes one of the greatest differences between survey data and administrative data.

In terms of process, the traditional survey framework enables us to:

- control the observation protocol: it may be complex for certain types of studies, for example multimodal surveys or surveys involving diaries²¹, or else surveys outside the home for which there is no sampling frame,
- verify, where relevant, data at the source (interaction between interviewer and respondent, or clerk and respondent),²²
- estimate error thanks to our knowledge of the sampling frame, and the random sampling process, but also taking account of imputation (or reweighting) models for item or unit nonresponse.

20 For example, structural surveys of businesses conducted in a given year Y often focus on figures for the year Y-1, essentially looking at the most recent accounting year. Annual turnover cannot be known until the end of the year, and this information can therefore only be recorded the following year.

21 In all Household Budget surveys, data collection involves a combination of one or more interviews, and diaries or logs maintained by households and/or individuals, generally on a daily basis.

22 The work of a survey clerk comprises various responsibilities, including manual data checking (micro or macro), monitoring, recontacting non-respondents etc. Survey clerks have an essential role to play in ensuring the quality of official statistics.

1.e. Representing the process: the GSBPM

Official statistics are presented in a standardised manner, following a shared, international framework for the process of producing official statistics: the Generic Statistical Business Process Model, or GSBPM (UNECE 2019).

The GSBPM sets out the steps of the process, which are divided into eight phases, and thus serves as a reference model. It is designed to be applied as statistical institutes see fit, in the manner most appropriate to their organisational structure and constraints (Erikson 2020).

Those eight phases are as follows:²³

- 1. Specify needs**
- 2. Design**
Including descriptions of the variables, sample, collection process, outcomes and production system
- 3. Build**
Including the collection instrument and the components of the process, plus testing of this process
- 4. Collect**
Creation of a sample selection framework, organisation, execution and finalisation of data collection
- 5. Process**
Including integration, coding, data editing, imputation, calculation of derived variables, finalisation of the data file
6. Analyse
7. Disseminate
8. Evaluate

Focusing exclusively on the first 4 phases, before data processing begins, in the case of surveys the process of statistical production consists of overcoming various difficulties: converging towards a purpose (and thus variables and classifications), getting to grips with the underlying theoretical models and conventions specific to the domain, formulating questions which are consistent with the goals identified, selecting the most appropriate sampling frame, selecting a sample designed to achieve the required precision, building an effective data collection tool, developing a collection protocol suited to the different sub-populations and configurations, dealing with nonresponses (partial or total), coding responses using the relevant statistical classifications etc. Mastery of this process as a whole enables us to construct all of the metadata (Sundgren 1993, Bonnans 2019) required for each successive step.

With administrative data we encounter problems and difficulties of an entirely different nature, calling into question the validity of the sequence of phases defined in the GSBPM, at the very least in the first phases preceding data processing, or Phase 5.

²³ Highlighted in bold are the steps most relevant to this working paper.

2. Using administrative data to produce statistics

The opportunity to use vast sources of data (including private data, indeed) may at first sight appear to be a potential goldmine for official statisticians: a form of exhaustiveness, allowing them to reach smaller sub-populations, with no sampling variance, significant cost reductions etc. But Hand (2018) effectively demonstrates that these advantages are not guaranteed, and it would be **premature to conclude that the lack of data collection costs alone is sufficient to justify the use of administrative sources**, even in the absence of a broadly agreed-upon theoretical framework.

First of all, it is essential to have a clearer understanding of what we mean when we talk about administrative data, beginning by identifying the principal characteristics of such data, before considering their uses. But there is still more work to be done: adopting a more rigorous approach to this issue, we must begin by defining the concept of data, and its constituent dimensions. Having established this conceptual framework, we can begin to examine the characteristics of administrative data when *used for statistical purposes*, a framework which will prove useful in the ensuing chapters.

2.a. Some characteristics of administrative data

The first dimension which comes to mind immediately is the **institutional** nature of administrative data: when discussing such data, we explicitly cite their origin and the nature of their producer, more often than not a governmental organisation.²⁴ The OECD defines administrative data as “*the set of units and data derived from an administrative source,*” with the latter defined as “*the organisational unit responsible for implementing an administrative regulation (or group of regulations), for which the corresponding register of units and the transactions are viewed as a source of statistical data.*” (OECD 2008) These are data of an official nature, built upon legal foundations, and their very existence may entail certain rights or responsibilities, with very real consequences for people’s day-to-day lives: paying taxes or fines, receiving pensions, qualifying for subsidies etc. Unlike private data,²⁵ administrative data are often associated with obligations (for example paying social security contributions), and they do not confer commercial or strategic benefits upon the administrative agencies that hold them.

In the existing academic literature, numerous authors have placed special emphasis on the **use** of administrative data. Hand (2018) defines them simply as “*data generated during the course of some operation, and then retained in a database.*” Alain Desrosières offers a slightly more precise definition: “*An administrative source is derived from an institution whose purpose is not to produce such information, but whose regular activities require them to keep, in accordance with the general regulations, individual files or records, whose aggregates are simply by-products, whereas the individual information remains the central element, particularly for the individuals or businesses concerned.*” (Desrosières 2004)

Fundamentally, what sets administrative data apart from survey data is their **purpose**: they are created to support administrative operations, with the intention of facilitating the implementation of an administrative process, whereas the intention of a survey is to observe. For the administrative bodies involved, establishing databases is not an objective in its own right; databases are supporting materials, used to manage or support operational processes. The data are thus structured according to their domain-specific codes and concepts, associated with the processes they support and their accompanying constraints: in the example of hospital data they include identifying information for

²⁴ In this paper we use various terms, more or less interchangeably, to refer to such organisations: “administrative agency”, “administrative entity”, “public body” or “public sector body”.

²⁵ This includes data originating with businesses in the private sector, such as scanner data (Leclair 2019).

medical professionals and care providers, patient care pathways, and classifications of patients into groups to facilitate reimbursement by the public health agency.

As a result, an administrative data item is not necessarily an observation: it may correspond to an administrative decision, for example the allocation of a social security benefit.

Example: for the public health agency, a simple appointment for an eyesight check-up with an ophthalmologist may be recorded in different ways in the information system: as a single ophthalmological consultation, or else as a series of medical procedures (all actions performed by the doctor during the consultation). Since the sums payable in these two cases are not the same, doctors may decide to optimise the codes they report in order to maximise their remuneration. Doctors in Sector 2²⁶ may tend to declare a single consultation, for example, whereas those in Sector 1 may be more inclined to declare individual acts.

Such optimisation decisions may also vary over time, and in different parts of the organisation. This is all the more true when the management of administrative processes is decentralised.

On a more general level, administrative data can be regarded as “action triggers” rather than passive information. They are constructed for a specific use (for example: calculating the remuneration of doctors, allocating social benefits, levying taxes correctly etc.), and are not primarily designed for observation purposes: Hand defines them as **by-products of operational processes**, or even “residue”. He describes “*data exhaust*”²⁷ as “*that which is left over after the organizational machinery has used the data to drive itself forward.*”

Generally speaking, these data are not primarily intended for **statistical use**, unlike survey data which are specifically designed for that purpose.²⁸ With administrative data, statistical use represents a secondary use, a form of reuse. Nonetheless, some of these data sources (annual declarations of social security and tax data, for example) have been used in the production of statistics for many years (INSEE has been using tax data since 1950), an arrangement facilitated, in France, by the fact that official statistical services have a guaranteed legal right to access these data.²⁹

Nevertheless, administrative data also have one major drawback: there is no **guarantee of their continued availability**. Dependent as they are on the processes they support, administrative data are liable to evolve significantly in terms of their scope, their concepts, and even their continued existence. Recent changes to French tax laws (the abolition of council tax – the *taxe d’habitation*³⁰ – on primary residences) have led to the disappearance of the data associated with the scrapped taxes, data which were of great importance to official statisticians: INSEE used this information to define households within its survey sampling frame. The abolition of the council tax (Bach et al. 2023) was thus a great loss to official statisticians, a loss about which they were not consulted, as the link between individual citizens and their homes disappeared from one day to the next.

Even if a source does not disappear altogether, variables which are of central importance to statistical processes may be removed or modified with little warning.

26 A sector 1 or 2 doctor is a practitioner whose fees are regulated by the French social security system. In sector 1, consultations are reimbursed at the agreed rate, while in sector 2, additional fees may apply.

27 A term which evokes both the depletion of a resource and the gases produced by car engines: that which remains of the raw material data once it has been extensively employed for its primary purpose.

28 As a result, not all data are useful for statistical purposes, far from it, and there are many forms of what we might call “secondary” data.

29 Article 7 bis of Act [51-711 of 7 June 1951 on statistical obligations, coordination and confidentiality](#).

30 The various sources cited in this document are described briefly in Annex 2.

Example: tax exemptions on overtime hours may mean that they are no longer declared as part of net taxable salaries... a fundamental variable for estimates of wage variation.³¹

As a general rule, the fact that these data do not fall under the responsibility of the statistical system naturally means that there is a **lack of control**: they are not produced under the same conditions as survey data. In the next section, we shall explain some of the differences in detail, starting with a formal definition of the different dimensions of data as a concept, from a statistical perspective.

2.b. The concept of data... and a new analytical framework

There are many definitions of data in the existing literature, for example (Kitchin 2014), (Borgman 2015), or (Caron et al 2020). For our purposes we will use the more operational definition proposed by Thomas Redman (1997), itself based upon a definition which is well-established in the realm of information science.

Data can be defined as constituting a triptych (concept, domain, value).³²

The **concept** represents the meaning of the data, which can be described by a label: for example, “the height of the Eiffel Tower”, “the gender of a person”, “the start date of a contract”, “the principal activity of a business”. It can be understood as an attribute of an **object**, or, to use a more familiar term, as a **variable** associated with an object. In these examples, the objects are the Eiffel Tower, a person (a specific individual), a contract (a specific contract), a business (a specific business). The attributes, or variables, are: height, gender, start date, principal activity.

The **domain** is the range of possible values which might be associated with the concept, and their interpretation. Let us say that height, for example, must be a positive integer expressed in metres. Gender, meanwhile, must be one option from the range {0, 1}, themselves associated with explanations (for example 0 = male, 1 = female), or {0, 1, 2} with an “other” category. As for date, we must decide which calendar format will be used, incorporating the appropriate constraints into the definition of the domain (e.g. the day must be less than or equal to 31). It is worth noting here that *format* is simply a manner of representing the value domain. For example, there are various date formats in existence (DDMMYY, DD/MM/YYYY etc.), but the choice of format changes nothing about the extent of the domain.

And the **value** is the specific value assigned to the variable for a given object. The thing we tend to conflate with the term “data”.

Concept and domain are also integral to the notion of data, a point of fundamental importance here: “324”, for example, is not a piece of data; it is simply a number. It may be associated with a concept such as “height of the Eiffel Tower”, “number of residents in a village”, or “weight of a specific motorbike”, in each case becoming a different kind of data, and we need to determine the characteristics of the domain, for example by specifying the unit used to measure height (feet? metres?).

If we have statistical purposes in mind, and recall the characteristics proposed in the section on survey data, two other dimensions of data as a concept have important roles to play.

The first is **scope**, which serves to restrict the spectrum of possibilities for the object in question: for example, an administrative measure which applies only to businesses with fewer than 20 employees founded after a certain date, or transactions of a certain kind completed in a given year, or pensions

31 In practice, the tax declaration instructions change over time: overtime had to be declared before 2020, was then exempted between 2021 and 2023, and has been reinstated since 2024.

32 On this subject, see also (Rivière 2020).

paid in a given period, up to a certain threshold, to persons with certain characteristics (age, number of years in employment etc.).

The second is **time frame**: data are often characterised by a *reference date (or period)* (for example, a person missed work through illness on a given day), which is itself a very useful piece of metadata (Gartner 2016).³³ But this reference date is not identical to the *date (or period) of acquisition*, i.e. the date on which the information was gathered: for example, a business survey conducted in 2024 concerning the 2023 financial year.

In order to better get to grips with administrative data, we make use of an analytical framework comparable to that applied to survey data, and also comprising 5 dimensions (Froeschl & Grossmann 2000):³⁴

1. Object
2. Scope
3. Variable
4. Value domain
5. Time frames

Figure 2 – Framework for the analysis of administrative data

This framework does, however, differ slightly from that applied to surveys: we no longer talk in terms of observation units, but of objects;³⁵ variables are attributes of objects; the value domain does not always refer to a system of classification; the notion of scope is similar to that which applies to surveys, except that here it is specific to the requirements of the administration; finally, the time frames are, in most cases, very different from those employed in surveys.

2.c. Administrative data for statistical purposes

For each of the five dimensions proposed above (object, variable, domain, scope, time frame), we shall now identify the specificities of administrative data and compare them with survey data. Much like (De Broe et al 2020), this allows us to compare the advantages and disadvantages of the two approaches.

- **Object**

The goal of administrative agencies is to execute their operational processes successfully, delivering services for users, collecting taxes, granting rights, paying benefits etc. and recording the corresponding events. The objects involved, which sometimes appear to be very simple, may prove to be complex on account of the multiple connections between them: in the hospital sector, for example, patients' care pathways involve multiple "objects" in addition to the patients themselves, each of which has its own dynamics and attributes: doctors, prescriptions, diagnoses, establishments etc. These are "administrative" objects, as distinct from the objects of statistical "studies".

Objects must be created in the information system (new patient, new doctor, new establishment). This may take the form of simply declaring an event, such as a birth, a physical event resulting in a

³³ Metadata can be defined as "data which provides information on other data."

³⁴ These authors make a similar distinction, as their work straddles the statistical and administrative realms: *populations and populations units, variables and value sets, coverage, data production dynamics*.

³⁵ Objects specific to the administrative entity, but which need not necessarily be associated with an observation process.

new entry in the register of individuals. For businesses, the corresponding event would be the creation of the business “object” in the information system (or register), officially marking its establishment.

Administrative data often refer to individuals,³⁶ which may give the illusion of simplicity: as supporting materials for administrative processes, data actually refer to the **capacities** of individuals : a person may interact with the administration in different capacities.

Examples: a natural person may be considered – depending on the administrative domain, practices and persons involved – a service user, a voter, a taxpayer, a beneficiary, a student, a healthcare professional, a guarantor and so on. In each case, the analytical angle and variables are different.

As such, the link between administrative object and statistical unit is not necessarily bijective.

A single natural person may correspond to several administrative objects (for example, where the same person is registered as a taxpayer in several administrative areas). Conversely, a single administrative object may cover several natural persons, as in the case of a benefit claim involving a claimant and one or more dependants.

The multiplicity of objects involved in administrative processes also opens up a variety of statistical possibilities.

We might cite the numerous potential uses of land registry data which, depending on the source, may relate to different objects: homeowners, homes, transactions etc. (André, Meslin 2022). Used via Résil³⁷ (Lefebvre 2024) or Fidéli³⁸ (Lamarche, Lollivier 2021) to determine the location of homes, these data may also be used to estimate property wealth (André, Arnold, Meslin 2021).

With surveys, however, in most cases the context is simple: the “object” is the individual / household / business surveyed, an entry “created” in the information system by means of the sampling frame and the selection of the sample.³⁹ The issue of roles and connections between objects is thus less problematic, although it does exist (child–parent, business–establishment).

- **Scope**

Administrative processes often concern only a subset of the population: for example, all individuals *liable for a specific tax*, or all businesses *eligible for a particular subsidy*.

Scope is a matter of “intension”, defined by its characteristics: all objects displaying a certain property, for example, all establishments with fewer than 10 employees. This scope, simply expressed, allows us to deduce an “extensional” list of objects, which requires us to know the total population: for example, the SIRENE⁴⁰ register for businesses, or the National Register for the Identification of Natural Persons for individuals⁴¹ (Espinasse, Roux 2022). In administrative contexts, the definition of the scope may evolve over time, quite simply because the legislation

36 Of course, we also have administrative data concerning businesses, establishments, administrative agencies etc.

37 Répertoire statistique des individus et des logements (Statistical Register of Individuals and Dwellings), built on several administrative data sources

38 Fichier Démographique sur les Logements et les Individus (Demographic File on Dwellings and Individuals, FIDÉLI) file generated through the management of housing tax

39 In most cases, but not all: we may find ourselves dealing with more complex objects such as business units (Moreau 2024).

40 Système Informatique pour le Répertoire des ENtreprises et des Établissements (Computer system for the business and establishment register)

changes (for example, a measure might apply to all businesses with fewer than 10 employees, until a reform makes it applicable to those with fewer than 5).

Example: the Individual Social Declaration (French: *Déclaration Sociale Nominative*, or DSN) is applicable to all employers, both public and private, who must declare salary information for all of their employees (Humbert-Bottin 2018). Its scope has evolved in recent years, since it only applied to the private sector until 2021.

Yet, as we have seen, processes may encompass multiple objects, and scope may also be defined with reference to these different objects. With regard to managerial objectives, the scope of an administrative process may thus differ from the scope of statistical interest, even if the objects concerned are of the same nature.

Examples: the scope of the GMBI⁴² (*Gérer mes biens immobiliers* – Manage my properties) data source encompasses all forms of premises, but the statistical uses envisaged are limited to residential buildings (housing).⁴³

Last but not least, the intended scope may not necessarily be entirely covered by the process. Coverage may be incomplete for several reasons: specific exemptions, missing declarations from the individuals concerned etc.

- **Variable**

This disparity between “administrative” variables and “statistics” is probably the most familiar and well-established topic of discussion: we know that there is no reason why administrative variables should meet statistical needs precisely, because there is no guarantee that their objectives are the same.

Example 1: income, not to be conflated with the administrative variable “taxable income”. By way of an illustration, the wages earned by interns are not taxable, and are thus not counted as household income by the tax authorities.

Example 2: in the DSN, the distinction between different forms of remuneration does not satisfy all statistical needs. It is not possible to distinguish between regular and one-off bonuses, or payment for periods spent on call.

The definition of a variable may vary from one source to the next, which makes it difficult to compare and use them for statistical purposes. For example, when building a boat for export, down payments made in the years preceding its delivery may be counted as export turnover by the public finance directorate (DGFIP). The customs authorities, meanwhile, record the price of the boat when it leaves French soil as the total value for turnover from the export sale.

- **Domain**

The domain represents the spectrum of values regarded as admissible for a given variable, which must be capable of being defined formally: *domain definition* (Batini, Scannapieco 2016).⁴⁴ This brings to mind questions of intervals, quantitative values, lists of values, selections from a long list of existing values etc.

41 Répertoire national d'immatriculation des personnes physiques (National Register for the Identification of Individuals, RNIPP)

42 See Desforges (2021), p. 1110.

43 The fact that the same word is used does not necessarily mean that the concepts are strictly equivalent.

44 See pp. 24-25.

It must be clear that the values must be accompanied by labels or metadata defining their meaning. For example, it is not enough to say that for the variable “gender”, the list of possible values is {0, 1}; we need to assign a meaning to those values (0=male and 1=female, or vice versa?). In the case of quantitative data, we must indicate the unit of measurement: Euros or thousands of Euros? Residents or thousands of residents? Metres or kilometres?

Understanding the domain, for administrative data, also requires us to understand how missing values were handled (-1? 99? 9999? N/A?), something which is often poorly documented. We must also be wary of “default values”, i.e. the values we reach for when we do not quite know what to say (N.B. these are not to be confused with missing values).

In the GMBI, for example, the default answer for the variable “are these premises furnished” is “No”.

We must thus distinguish between genuine “No” responses (or some other default value) and nonresponses. Moreover, the instructions associated with default values are not always the same for everybody.⁴⁵

Finally, we often find that the “domains” associated with data are in fact classifications, which play an essential role in administrative activities: legal categories, establishments, activities, grades, etc. Bowker and Star (2000) also emphasise this point: “assigning things, people, or their actions to categories is a ubiquitous part of work in the modern, bureaucratic state.”

The challenge is to understand the characteristics of the domain properly (again, documentation is required...) and specify the differences with any similar statistical classifications. Administrative classifications serve operational purposes: for example, dividing a country into regions simply reflects the reality of its administrative organisation. Paradoxically, the fact that an administrative agency opts to use an “INSEE classification” does not necessarily facilitate matters: they may be working with an outdated version, or a version which is partially aggregated or transformed. Ostensible references to official statistics must thus be treated with caution.

- ***Time frames***

The dual time frame of statistical surveys is usually very clear: data are collected *within* a clearly defined and declared period (the *acquisition period*, or *collection period*), during which time we *become aware* of this information. That information *relates to* a period, or date (*period*, or *date*, of *reference*). The DDI⁴⁶ (Data Documentation Initiative) metadata standard, an international framework widely used in statistics, is clear about this distinction. For surveys, it is preferable for these dates and/or periods to be relatively close to one another, and uniformly applicable to all units.

With administrative data, the simple difference is that, in the majority of cases, there is no collection phase (in the sense that there is no universally applied dual time frame). One exception which proves this rule: *administrative declarations*, characterised by a *declaration period* and a *reference period* to which the declarations refer. The declaration process (for tax returns, social security declarations and the like) is thus similar to the data collection process in terms of its time frame.

⁴⁵ For example, in the information system operated by France’s public health agency, to ensure that patients without designated family doctors (because they cannot find one due to a lack of available doctors) are not penalised in terms of quality of care, the health insurance funds (CPAMs) record a fictitious code for designated doctor. The techniques used vary from one CPAM to the next, and are not necessarily known at national level.

⁴⁶ See https://ddialliance.org/ddi-codebook_v2.5. This corresponds to the tags <referencePeriod> or <timePeriodCovered> on the one hand, and <collectionDate> or <dataCollectionPeriod> on the other.

Moreover, administrative data are dependent upon external events which may occur at any moment: births, deaths, change of address, pension applications, requests for services etc. As such, there is absolutely no control over the *reference date*, and the *date of acquisition*, or its equivalent, is often entirely dependent on how promptly users report the event (submitting a doctor's note a little late is an easy mistake to make...).

It is worth noting that administrative events may also be events internal to the administration, for example the decision to allocate benefits or subsidies, the decision to issue a fine etc. Here, once again, the functional priorities of administrative management do not always extend to close control over dates, such as the date on which decision is made (e.g. legal rulings).

For statistical purposes, this lack of control over both time frames is problematic: there is no guarantee that we have an exhaustive picture of events, because of the potential delay in their declaration, and the risk that they will not be declared at all.

- **Summary**

Ultimately, if we wish to estimate the total described in the first section of this chapter, working with administrative data requires a very different algebraic expression than that employed for survey data, formulated as follows:

$$\hat{X}_{\tilde{C}_k}^t = \sum_{\substack{i \in \tilde{U} \\ i \in \tilde{C}_k}} \tilde{X}_i^{t_i} \quad (5)$$

- Where universe U, classification C and variable X, specific to the administrative entity, need not correspond to U, C and X.
- Where, for each unit i , the reference date/period t_i , is expected to be close to t , although the values of t_i may differ across units.
- Where, finally, none of the terms is derived from a predefined random model, which makes “traditional” calculation of sampling error impossible.⁴⁷ Nor is it possible to calculate nonresponse error in the traditional manner, due to a lack of knowledge regarding the sampling frame. However, other methods may be used to assess coverage shortcomings (capture-recapture methods).

Estimating total error (much like total survey error) is a complex undertaking, but not an impossible one. Reid et al. (2017) propose a framework which allows us to identify different types of error. Nonetheless, they remain difficult to model. Berkovsky et al. (2025) propose a total data quality paradigm which takes the specific challenges of administrative data into consideration, but extends beyond the probabilistic framework .

2.d. Statistics derived from administrative data vs statistics derived from surveys

Compared with survey data, administrative data offer numerous clear benefits for statisticians (De Broe et al 2020):

- No direct data-collection costs... for official statistical services
- No additional survey-response burden⁴⁸

⁴⁷ At a more general level, we are interested in total error: with surveys, this includes sampling error and also non-response error (which can be modelled using the information at our disposal regarding the sampling frame)... Nonetheless, things are not always entirely perfect with surveys: certain forms of error (such as measurement error) are hard to estimate.

⁴⁸ ... though it may be replaced by an administrative burden (the need to fill in more forms).

- More granular data, allowing for very fine analyses (geographical, sectoral, specific age groups)
- More opportunities for data linkage: if there is no shared identifying code (in the NIR⁴⁹ or SIREN⁵⁰ registers, for example), identifying elements (first name / surname / date of birth, or address / company name) are often identical, allowing us to pair and match files.

However, as we have seen, the status, purpose and life cycle (connected to their use) of administrative data differ substantially from those of survey data, and this has consequences for the properties of statistics derived from them. Returning to the 5 dimensions of our analytical framework, we can see that **official statisticians** are no longer in control of many of them.

- Scope, variables and domain are all decided by and for the administrative agency, usually with no consideration for statistical uses. The challenge is thus to determine the characteristics of these elements, allowing us to subsequently transform the data for statistical use. The quality of this work is also heavily dependent on the metadata attached to administrative sources, which are not always well documented, or may be inaccessible.⁵¹
- Data do not always pertain to the kind of objects which statisticians are used to dealing with (individual, household, business, establishment), but instead to objects defined for operational purposes (transactions, invoices, premises, electricity meters etc.).
- We often have no control over the homogeneity of the dual time frame (reference / acquisition time), with the noteworthy exception of administrative data based upon declarations from fixed declaration periods.

More broadly, the data-generation process lies outside the statistical office's control, to the extent that statisticians are unable to verify the data with their producer, since the link to the original data subject (the declarant) has been broken.⁵² When we are not responsible for establishing the probabilistic framework (survey plan, imputation models etc.) in advance, we lose the capacity to measure error using this framework. Nonetheless, this does not prevent us from defining other measures *a posteriori*. This is an important point.

2.e. Consequences for the process, compared with surveys

Preamble: the aim of this section is to draw parallels between the survey process and the process used to transform administrative data for statistical use, with reference to the GSBPM. It is worth noting that several projects have been launched at European level to explore the possibility of adapting the GSBPM to administrative data, at least in some areas. Eurostat has made several attempts to align the steps of the GSBPM with the process applied to administrative data (Eurostat 2017c⁵³).

The first steps in the GSBPM process are: specify needs, design, build and collect.

With administrative data, collection in the traditional sense is not required. There is no need to select the sample, design the collection protocol and tools, or actually collect the data. Nonetheless,

49 NIR is the national insurance number, and the identification number in the National Register for the Identification of Individuals

50 SIREN is the identification number in the business register SIRENE.

51 For example, the confidentiality applicable to tax matters sometimes prevents us from attaining a clear view of the processes used to calculate certain characteristics.

52 This includes the very existence of the data, as there is no guarantee that an administrative source will remain available *ad vitam aeternam*. This has consequences for the long-term viability / comparability of the statistics.

53 Indeed, Part 2 is entitled: "Applying the GSBPM to the usages characterised by integrated administrative data". Part 3 deals with the processing phase.

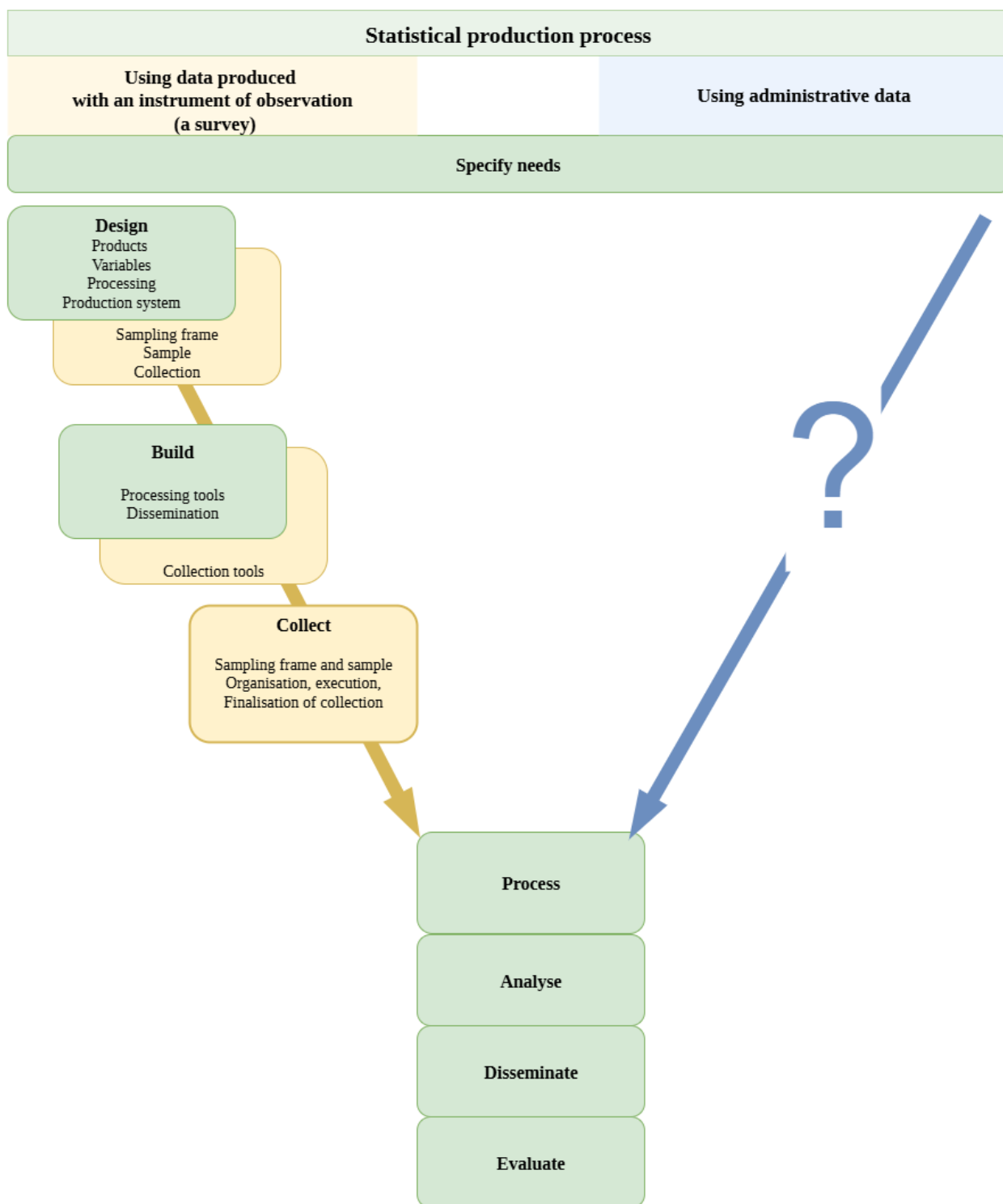
work is still required to specify needs (within the constraints imposed by the content of the administrative data), and design and construct some things.

Joining together these two possibilities – surveys or acquisition of administrative sources – we have a statistical production process in a **Y** shape:

- The upper left branch of the Y corresponds to the first 4 steps of the GSBPM,
- The trunk of the Y represents the next steps in the process, from “processing” onwards,
- The top right branch of the Y corresponds, for administrative sources, to the process preceding statistical processing, as explained in the present document.

For official statisticians, the design and development process consists of aligning parameters with “their” usual objects: scope, variables, domains, time frames. A process must be developed for the purpose of *packaging* the data received from administrative agencies, transmitted in files over which statisticians have little or no control, and which must be taken as is. In the words of a recent article, these are “*it is what it is*” datasets (Lothian, Holmberg, Seyb 2019).

This new process is applied not to isolated data units, but to a broader set of data: an “**administrative source**”. But do we have a shared vision of what we mean by this term, a widely accepted definition of the concept? As we shall see in the following chapter, there is no self-evident or universally accepted definition..



This diagram represents the steps of the GSBPM, highlighting some of the sub-phases. We distinguish between those sub-phases connected to the desired statistics (the goal) and those sub-phases more concerned with how to achieve this goal. In the latter category, we highlight the steps relating to the design, construction and implementation of an observation instrument in a survey context (in yellow in the diagram). These are the two sub-steps whose workings we shall attempt to describe in detail, in the specific context of the usage of administrative data for statistical ends.

3. Administrative information systems and administrative sources

Within administrative agencies, data do not just emerge from a collection process and end up being stored somewhere. On the contrary, they exist within an *information system*, which in some respects might be considered to constitute their natural habitat. This is a complex environment in which data are constantly evolving, in response to data management actions. The concept of the “source” is thus largely a construct, invented by statisticians for their own purposes.

In this section, we offer a brief description of what actually constitutes an administrative information system (IS), before addressing the question of how data can be “frozen” in time, and then attempting to define administrative sources as being derived from information systems. As we shall see, switching between worlds, i.e. moving from one domain (in this case, administration) to another (statistics), is a delicate operation which requires us to adopt a multi-step process.

3.a. The concept of the administrative information system

An information system (IS) is an organised set of resources (hardware, software, people, procedures) making it possible to collect, classify, process, store, disseminate and provide access to information within an organisation. Its purpose is to support and improve operations, decision-making and the control and coordination of activities, with a view to attaining the strategic and operational objectives of the organisation (Kalika et al 2023).

Within an administrative agency or business, the general information system usually comprises multiple sub-systems (e.g. human resources, purchasing, customer relations etc.), each with their own objectives and operational processes (for example, recruitment processes in HR information systems), which are often interconnected.

Information systems are worlds of their own: complex, overlapping, ever-changing and varied as a result of being constructed piecemeal over time. They necessarily involve a certain degree of technicality, domain-specific concepts and terminology, organisational cultures and specific objectives.... and they naturally generate and utilise a large quantity of data, for their own purposes. These data are embedded in the IS and are largely unfit for research, analytical and statistical purposes, not least because they are constantly evolving in response to the way that administrative processes are managed. The data exist to support specialised work processes.

It is thus important to bear in mind that administrative data are not, in most cases, fixed and immutable. On the contrary, they are live data subject to continuous change in a variety of guises (addition and removal of objects, modification of attributes, events involving one or several objects etc.).⁵⁴ Data exist within *databases*, structures designed to facilitate this process of permanent evolution without compromising the integrity of the whole (Elmasri, Navathe 2016). As such, raw data cannot be used directly for statistical purposes until they have been transformed: because of their ever-changing status, each new aggregation operation will produce a different result that is relevant only to the precise moment in time at which it was produced.⁵⁵

The information systems we are referring to here belong to *public entities*. It is these entities that transmit what we call administrative “sources”, via the operational units entrusted with this task. Generally speaking, this is not an exercise in data collection, and the data are not necessarily

54 We must always make the distinction between developments at “instance” level (adding or removing an individual, for example) and developments at the level of the data model, which is to say at “definition” level (adding the notion of a ‘beneficiary’, for example).

55 Suffice it to say that there is always some form of sub-system devoted to steering operations, which contains frozen data and thus partially responds to the preoccupations of the statisticians.

observations: official statisticians may sometimes have a tendency to project their own representations onto data, and wrongly so.

An administrative IS is a vast, protean and often evolving entity, and cannot, as such, constitute a statistical data source. Moreover, it will almost always remain unknown to the statisticians. For them to do their work, they need a file containing frozen data. In order to produce statistics, once the key features of the IS have been assimilated, they need a clear definition of the methods used to **freeze** the relevant datasets within the IS.

Conversely, when conducting surveys, we put in place a time-bounded observation process, thus yielding by construction a fixed dataset, which more often than not contains a unique, dated value for each variable of interest.

But how do we freeze something which is constantly moving? And where do we start?

3.b. Freezing data: a necessary construct implying some decisive choices

When producing statistics, we cannot take as our starting point data whose values are liable to change at any moment. One way or another, we need to get our hands on a file containing fixed, unmoving data. It is therefore crucial to be able to take a “snapshot” of our ever-shifting data, “**freezing**” them at a moment in time. It is not necessarily the task of official statisticians to take this snapshot, but they must understand *how* it was taken, as the process always involves several choices.⁵⁶

We have identified four potential freezing scenarios⁵⁷ which depend on the structure of the IS in question, ranked here from the simplest to the most complex:

- ***When the administrative source is based upon administrative declarations from within a controlled collection period***

In a declaratory process (social security declarations, for example), data are naturally “frozen” in the sense that the process – by construction, independently of the wishes of statisticians – yields data which do not change: they reflect what was declared at a given moment in time T.⁵⁸

A natural by-product of this process is a file containing all of the declarations and declarants for the collection period, as defined in the collection process. In a sense this might be regarded as the “**canonical**” **freezing** process, repeated for each declaration period, aligned with the organisation’s needs and impervious to other considerations.

In practice, of course, the snapshot is never perfect; between one collection period and the next, retrospective modifications may be necessary (this is true of the DSN); nonetheless, these changes are of marginal importance, and for the most part the value recorded for each period is unique. Declared data are, by their very nature, frozen and dependent upon the declaration process.

⁵⁶ Much like the choices we make when taking an actual snapshot: subject, date, place, lighting, framing, focus etc.

⁵⁷ We use the term *freezing*, or *data freezing*, to describe the action of fixing data at a specific moment in time. Alternatively, we might have used the word “extraction”, but we feel that this latter term does not sufficiently convey the sense of freezing/immobilising the data at a point in time, with all that this implies. And it is a little bit more complicated than “data snapshotting”.

⁵⁸ Although we might also envisage corrective declarations or delayed earnings (e.g. bonuses paid at a later date, which may need to be attached to the relevant period of activity).

Example: in the monthly PASRAU (Pay as you earn – Other Income) data source, corresponding to declarations of “replacement” income sources such as unemployment benefits, pensions etc. (Berthelot 2020), we note that in order to encompass 100% of payments made in a given month, we must analyse 5 consecutive monthly instalments.

Another example: data for income tax are derived from an operation which runs to a precise schedule (the 2024 declaration period opened on 11 April and closed on 6 June 2024), with a clearly-defined population⁵⁹ and period of reference (the year 2023). A frozen dataset can be obtained from this source by “simply” aggregating all of the declarations submitted to the tax authorities during the data collection period.⁶⁰

Declaration processes, whose numerous components are defined *ex ante* and in rigorous fashion (reference period, target population, concepts etc.) thus constitute good examples (Rivière 2018) of cases where the **definition of freezing is (almost) entirely unambiguous**. By this we mean declarations made *at the initiative of the administration* (tax declarations, for example), complete with an organised collection process.

In such cases, it is the administrative entity that sends the “file” to the statistical office, handles the concatenation of the declarations, and assumes responsibility for them. Nevertheless, even in this ideal scenario, the reality is rarely so simple: data may be delivered in several instalments, each with its own reference date/period, scope, and may even involve multiple providers, adding extra layers of complication. A whole **delivery process** therefore needs to be put in place, with all of the conventions that entails.

- ***When the entity in question already regularly produces frozen datasets for its own purposes***

Government bodies and businesses alike very often need a clear overview of their operational indicators: headcount, purchasing, budget, output etc. Management databases are living entities which are not capable of providing such an overview as they stand. Organisations thus find themselves obliged to produce regular “status updates”, snapshots of their current situation derived from these live databases.

To some extent, this means that we are able to pick and choose from several off-the-shelf data products (Meierhofer, Stadelmann, Cielibak 2019), as long as an agreement is in place between the relevant parties. These products can make statisticians’ lives easier, subject to certain conditions: firstly, they must be aware of their existence, then they must obtain permission to use them (by means of a formal agreement), be able to access a bare minimum amount of documentation, and of course be confident that their content is actually useful. Data products⁶¹ come in a wide variety of forms, but there are two in particular which are very commonly encountered.

The first involves information from inventory-type reports calculated using management databases for a specific date or period: for example, a full list of employees as of 31st December, or, for a library, a list of all loans made over the course of a given year. In both of these examples, the process of freezing the data may appear relatively simple.

59 A definition (in French) can be found here: <https://www.economie.gouv.fr/particuliers/declaration-revenus-reponses-questions#>

60 Although this freezing operation may seem simple (which it isn’t, as it happens), it is worth noting that it does not guarantee that the resulting data will be suitable for statistical uses.

61 It should be acknowledged that *data products* are in fact a far more wide-ranging proposition, extending beyond data in and of itself: “A *data product* is defined as the application of a unique blend of skills from analytics, engineering & communication aiming at generating value from the data itself to provide benefit to another entity.” The concept of *data products* is often associated with the concept of *data contracts*, building upon the notion of *interface contracts* which has long been used in the context of exchange systems.

The process may also take on far more elaborate guises, however, derived not from “snapshots” but instead from complex calculations involving a vast array of data: one thinks immediately of **accounting statements** (profit and loss accounts, balance sheets), or the various types of activity **dashboards** including indicators for output, HR, financial performance and so on. In such cases, the organisation is effectively producing statistics for its own purposes, i.e. for operational reasons and in order to fulfil reporting obligations. This exercise goes far beyond simply freezing live data at a certain moment in time: the data are subjected to transformations, aggregations and elaborate calculations.

These constructed, “off-the-shelf” datasets are defined by the owners of the associated databases. They allow statisticians to retrieve data which are no longer shifting and changing, but they must also accept that they have no control over the way in which the data in question were frozen and transformed, processes which are not necessarily well documented.

At this juncture it is worth citing an example from beyond the domain of administrative data and involving private data, but which is nonetheless enlightening: during the Covid pandemic in 2020, telecoms firm Orange decided to provide free access to its data. INSEE thus used data which had already been processed by Orange (not microdata) to produce estimates of the population present on French territory (Tavernier 2020).

- ***When an administrative source is based upon event-specific administrative declarations***

The classic example of an administrative declaration is the system used to calculate our tax, with a declaration period and a reference period which are both set by the government. But many other forms of declaration do not fall within this category, as they are tied to specific events.

Obvious examples include matters recorded in the civil register, such as births (Brumberg, Dozor, Golombek 2012; Espinasse, Roux 2022), marriages and deaths (Coudin, Robert 2024), but we might also include events pertaining to businesses, such as a change of address or company name.

With such declarations, neither of the time parameters can be controlled externally. The reference date corresponds to the date at which the event occurs, which cannot be decided insofar as it is a “real world” event. As for the time frame in which the event is recorded, it often depends on the willingness of the declarant who may well choose to make their declaration with a considerable delay, for example when medical paperwork is submitted to the health insurance agency sometime after the fact.

As such, with event-linked declarations, even if we are able to access frozen data (the event in question is not liable to change; it has occurred once and for all), the fact that we cannot control the time parameters has serious consequences for our degree of control over the scope of our analysis: if somebody forgets to declare an event, then we cannot know about it (whereas if the process were determined and organised by an administrative entity, we would be able to detect the absence of a declaration that was “expected”).

Example: a business ceases trading. This event may be recorded on the SIRENE business register several months (or even years) after the company actually halts its operations. The legal procedures involved in liquidating a company can be extremely lengthy.

A similar example, this time involving deaths: the National Register for the Identification of Individuals (RNIPP) contains a number of individuals over 120 years old. In reality, if an individual dies overseas and the death certificate is never sent back to France, then that person’s status will never be changed to “deceased” in it.

As a result, a form of dissymmetry is frequently observed in the registers between births and deaths of individuals or businesses. In order to overcome such coverage issues, statisticians must create concepts such as “statistical deaths”, ensuring that their figures for statistical units that are still “living” in administrative terms do not include units that are considered as being dead (or economically inactive, in the case of businesses)

Above and beyond the problem of delayed declarations, we must always bear in mind that the purposes of administrative management differ from those of statistics, not least because the former is rarely concerned with establishing a coherent picture of a specific moment in time (and may indeed have little interest in coherency at all)⁶².

- ***Other situations in which freezing methods are not unequivocal***

In the final scenario, which is complementary to the three others, we are not in possession of administrative declarations, and there are no off-the-shelf data products suitable for the job. In such cases, we must construct a frozen dataset which is suited to the purposes of the statistical study, using management databases. This is probably the most common solution in cases where no declarations are available.

This is true of statistics pertaining to the justice system, for example, where we must convert “live operational data into date-marked statistical data,” (Chambaz 2018) but it is also true of the medical-social sphere. (Berthe 2025)

In such cases, defining the methods to be used to freeze the data, and even the methods used for transformation purposes, is a matter of **cooperation** between data providers and statisticians. For the statisticians, the first priority is to gain an understanding of the existing data in the databases, their significance, life cycle, and how they are used within the providers’ specialised processes. Only then can they identify the most relevant data, and define the terms of an agreement between the two entities. In France, numerous agreements have been established between the official statistical services (INSEE and MSOs) and administrative agencies.

The text of the agreement must explicitly state the variables involved, the relevant scope, the temporal dimension and the transformations involved. This allows the creation of data products which are not off-the-shelf, but **tailor-made**, built to specifications jointly established by official statisticians and data providers. This means that we are no longer wholly dependent on the systems that administrative agencies have seen fit to build... but the preliminary work involved is a much longer process: the resulting **agreement**⁶³ forms the contractual framework for future exchanges, establishing a formal description which is useful for comparisons over time.

Drawing up such agreements can be a lengthy process, as evidenced by past experiences involving private data such as scanner data, which took 10 years to obtain (Leclair 2019), and data from bank accounts (Bonnet, Loisel 2024).

⁶² For example, when an employee is absent on sick leave, the time they miss will not be counted among the working hours for the month in question. However, if they nonetheless receive their wages in full, any subsequent corrections will not be recorded for the month in question (wages actually paid), but instead for the month in which charges are paid. Hence there is a potential time lag between hours worked and wages received, which poses a problem when calculating hourly wages.

⁶³ This subject will be further explored in Chapter 4, Section 4.b.

Working with private data: a few remarks

We have discussed the use of data from both public and private sources, but there are some noteworthy differences between the two which cannot be explored in detail here (see Lesur 2025). Using private data raises specific questions of confidentiality, and can entail significant costs (businesses keen to sell their data, cf. the case of mobile phone data), requiring special organisation (Chaleix, Mikol 2024).

Furthermore, the risk that a source will cease to exist, or that data will not be comparable over time (because data structures are liable to change) is probably greater when working with private data as compared to public data. The latter certainly do not come with firm guarantees attached, but nonetheless do offer a bit more stability. The notion of the declaration, meanwhile, with everything it implies in terms of formal obligation, is by and large confined to administrative data.

3.c. The “administrative source”: a fragile concept

In order to produce statistics, we need to be able to make the transition from a live information system to a structured data file, associated with a specific reference period/date and a clearly defined target population. In practice, however, information systems are often complex, overlapping, moving and partially redundant, which makes the task of freezing data more complex. Even those processes which appear to most closely resemble those used in surveys, for example processes based on administrative declarations, will often combine declared data with older, descriptive data pertaining to the target population.

Example: the individual income tax file provided each year by the Directorate General for Public Finances. In addition to details of all individuals who submitted tax declarations in the previous year (statistical field of interest), the file also contains details of all individuals who submitted declarations within the past 3 years.

Another example: applications for EU subsidies submitted by farmers. CAP (Common Agricultural Policy) data contain both declarations for the period in question and structural data on agricultural holdings which may come from earlier records (meaning that they are not always up to date).

As such, the act of freezing data often ends up including, even in the simplest cases, filtering information whose scope may not be limited to the declarants in which we are interested.

Generally speaking, the freezing process is used to derive a “source” from an IS. However, there is no single, unified process (there are a number of possible approaches), nor any guarantee of permanence over time (the structure of data within the IS evolves). One major consequence of this is that it is entirely possible to derive **multiple sources from the same information system**.⁶⁴ Freezing operations may be conducted at different moments in time, applying filters and selecting, aggregating and combining data. Freezing may come to resemble a “projection”, immortalising a live, ever-changing IS in a dated, often partial snapshot.

⁶⁴ It is also possible for a single source to be derived from multiple information systems.

Since many forms of projection are possible, sources derived from the same information system may differ on many points:

- scope (selection of objects and information),
- frequency with which the source is made available,
- timeliness of information,
- periodicity of information,
- period of reference,
- granularity of information,
- transformations performed on the data.

The notion of transformation⁶⁵ here spans a fairly broad range of meanings, including among other things changes to reference “objects”. Unlike surveys, which generally focus on a relatively small number of objects (often just one object: individuals, or businesses; or else two connected objects such as households and their component individuals), administrative information systems usually encompass several dimensions and multiple objects. In the DSN, for example, objects include businesses, employees, contracts, payments, workplaces, tax bases etc. For the land registry, objects include owners, premises, transactions etc. The multiplicity of these objects, and the non-unique nature of the connections between them (for example, an employee may have several contracts), means that the data freezing process may turn out to be complex and somewhat equivocal.⁶⁶

We use the term “administrative data source” to denote a fixed (or frozen) dataset pertaining to a specific moment or period of time, presented in the form of a file or files derived from an administrative information system by means of a more or less explicit and controlled set of selection, transformation and filtering rules.

The existence of multiple sources derived from the same IS raises other problems, and makes the notion of a “source” somewhat unstable. We may thus find instances of redundancy in the data, but also divergences: when subjected to different freezing processes, information which should theoretically be identical in both sources may turn out to be different. As a result, when official statistics agencies use multiple sources derived from a single IS, they must justify this choice and evaluate the costs and benefits of the different options available. “Sources” are not naturally occurring phenomena;⁶⁷ they are constructions resulting from multiple operations.⁶⁸

All of which is testament to the great difficulties inherent to putting together a catalogue of administrative sources. The Dataactivist website, which promotes a general culture of data awareness in society, highlights these difficulties in an article headed “The Ecosystem of public data sources,”⁶⁹ insisting upon the importance of distinguishing between sources on the basis of the entity that produced them, and whether they are primary or secondary sources. It also criticises the complexity of the “acronym jungle”.

In fact, even naming data sources is not a straightforward exercise: knowing the administrative IS from which it originated, or even the general name of the source, is often not sufficient to avoid confusion.

⁶⁵ We use the term “transformation” here to refer to the transition from IS data to frozen administrative data. We will subsequently use the word again to denote the transition from frozen administrative data to raw statistical data.

⁶⁶ Which also goes to show that, strictly speaking, “canonical” freezing for administrative declarations and similar cases is only possible subject to certain conditions.

⁶⁷ We might even wonder about the suitability of this term, widely used though it is...

⁶⁸ See (Denis 2018), and this interview with the author: <https://www.nonfiction.fr/article-9517-entretien-avec-jerome-denis-a-propos-de-son-livre-le-travail-invisible.htm>

⁶⁹ <https://open.dataactivist.coop/docs/culture-generale-donnees-section-4>

Example: the “POTE” data source produced by France’s Directorate General for Public Finances using income tax declarations is in fact broken down into four different sources, all used by the official statistics system, some of which relate only to a sample of the data, while others are pre-aggregated by the administration.

When seeking to define and develop an institute’s statistical processes, we must carefully describe the sources used, so as to determine whether we are dealing with a single source or multiple sources, and to date them accurately. Above all, the very concept of a source is inextricably linked with the proposed use of the data that it contains.

The official statistical system uses various data sources derived from the CNAF (National Family Benefits Fund) information system. These sources cover different reference periods and consolidation dates: annual figures, consolidated 6 months or 2 months after the end of year Y-1, for the entire scope; or else monthly figures, consolidated 6 months after the end of month M-1, for the entire scope, or else for a sample selected by INSEE for a DREES panel.

Extracting and sharing data incurs costs for the administrative departments responsible for producing them. They are also liable to change over time.

The CNAF, for example, has changed its source publication policy several times, first by abandoning consolidated results at M+2 months, and subsequently by reverting to a single monthly data transfer to INSEE, with the latter now responsible for sharing those data with the various official statistical services.

All things considered, the data source is a difficult concept to define. Depending on the circumstances in which they are “born”, and the methods used to freeze the data, sources may take a wide variety of forms. **Annex 1** contains an overview of the various classification schemes for administrative sources found in the literature, and our proposal for a new classification.

3.d. The delicate transition from one domain to another

Moving data from an administrative environment into the context of official statistics may appear to be a purely technical operation, a simple transfer of files.⁷⁰ In reality, there is far more going on: it is more like a change of universe or ecosystem, which is nowhere near as straightforward as one might imagine. In fact, this is a broad topic, and one which has generated a vast literature over the past fifteen years.

In a recent article, Borgman & Groth (2025) talk metaphorically of the “distance” between data creators and data users, exploring the many facets of this distance (methods, domains, objectives etc.).

- ***A stubborn cliché: it is important to share “raw data”***

The idea that certain types of data (administrative data, in particular) should be accessible to all, without serious restrictions, in order to promote transparency, innovation and socio-economic development, is a central tenet of the Open Data movement. A turning point in the history of Open Data was the summit held in Sebastopol, California, in December 2007.⁷¹ This conference brought together campaigners, researchers and practitioners with a view to establishing an initial framework

⁷⁰ ... which is not always as straightforward as we might hope: in *Informatique, normes et temps* (1999), Isabelle Boydens quotes an American general during the first Gulf War in 1990, bitterly ruing the fact that “it takes longer to transfer a database than a physical object”.

⁷¹ See (Goëta 2024) for a brief history of open data.

for open data, based upon eight fundamental principles: **data that are** complete, **primary**, timely, accessible, machine processable, non-discriminatory, non-proprietary and licence-free.

The second principle in this list reiterates the importance of making data available in their **primary** form, i.e. not transformed, taken directly from the source in detailed form. Tim Berners-Lee, one of the movement's foremost champions,⁷² insisted heavily upon this point in a February 2009 Ted talk which attracted much attention, and during which he encouraged the audience to chant "*Raw data now!*"⁷³

However, this stance was soon subject to some very robust challenges, not least in the collective tome *Raw Data is an Oxymoron* (Gitelman 2013). The authors who contributed to this collection question the very existence of "raw" data existing in a "pure" or "objective" form. Above and beyond the fact that the collection, structuring and interpretation of data are always contextualised by human choices, they warn against the illusion of neutrality: describing data as "raw" draws a veil over the various forms of bias implicit to the processes of data creation and analysis. Finally, and obviously, data always require metadata and interpretation if they are to be significant and useful.

In a book entitled *La fabrique des données brutes* ('*Manufacturing raw data*', 2014) Denis & Goëta also challenge the myth of raw data and the way it overlooks the transformations required to make data usable, reiterating the fact that data are always the product of human and technological labour. They go further still, coining the term "rawification" to describe the way data are treated: transformed into standard format, cleaned up and de-indexed in order to make them comprehensible and usable for third parties. They demonstrate that the process of "opening" data is not simply a matter of technical transfer; it also requires coordination between institutional stakeholders, allowing for technical adjustments and compromises.

As a concept, "raw data" has no absolute value: in the realm of business statistics, for example, official statistical services make use of accounting data, regarding these as "raw" data. But from the business perspective, these data are anything but raw: accounting data are in fact the output of a long process of construction, fine-tuning and calculation. Generally speaking, it would perhaps be more accurate to regard data as "ultra-transformed" products (see also Christen & Schnell (2024) on this subject). By way of an illustration, simply consider the sensors used to record temperatures, pollution etc.: the rawest possible form of data which they could produce would be a series of localised, dated, electrical pulses which are utterly unusable in and of themselves. Ultimately, and perhaps paradoxically, the "raw" data need to be *constructed* by packaging the data into a form consistent with their intended purpose.

⁷² The father of the world wide web, as inventor of the http protocol, and in more recent times a champion of the 'Semantic Web'.

⁷³ See https://www.youtube.com/watch?v=OM6XIICm_qo , around 11' – 11'30. Tim Berners-Lee touches upon many other subjects in this presentation, including the importance of standardising formats and his vision of a linked data system.

Transition and friction

In *A Vast Machine* (2010), Paul N. Edwards explores the central role of data in climate science, describing how they are collected, transformed and used to model and comprehend the global climate. He is particularly keen to make clear that climate data are not "raw facts", but instead "products" created by means of a process of interpretation and transformation ... and of course meteorologists and climatologists use different processes, since their objectives are not the same.

Meteorologists are primarily concerned with short-term forecasts, using very precise and localised data, whereas climatologists study long-term trends in climate systems, working with long-term averages and aggregates as they seek to better understand global interactions (in the atmosphere, oceans, cryosphere, etc.).

Edwards explains that these differences create different expectations with regard to the nature and use of data, which can lead to misunderstandings. Meteorologists thus produce observation data which are of great importance to climatologists, but which also pose various problems in the long term, not least on account of instrument bias (changing sensors, or the location of sensors) and discontinuities (observation networks which are not uniformly distributed in time and space, leaving sizeable blind spots in the climate archives).

As a result, climatologists are obliged to modify, supplement and homogenise meteorological data, a process which some meteorologists regard as tantamount to altering or manipulating the "raw facts".

In the same book, Edwards introduces the concept of "**data friction**", which has since been taken up by other authors (Bates 2017, Lehuède 2024). This term encapsulates the obstacles, tensions and efforts involved in collecting, harmonising, sharing and interpreting data in complex environments, which become even more complex at international level. The notion of "friction" is borrowed from both social sciences and physics, where it denotes resistance to movement.

Data friction arises at the interface⁷⁴ between two data "surfaces". Much like physical friction, it impedes the movement of data, increasing the cost and effort required (Lombardo 2023).⁷⁵ The movement of data may thus be restricted by mismatched formats and standards, methodological biases inherent to data collection, the limitations of technical infrastructure, or institutional tensions.

Courmont (2021) reports similar issues in a different context, this time outside the scientific field: Lyon City Council's decision to open up their public data to interested citizens, for example providing access to mobility data derived from administrative sources. Courmont observes that these administrative sources (such as data recorded by the highway maintenance department) are very often unfit for this new purpose, and must therefore be "detached" from their initial environment before being "reattached" to a different use. Once again, work is required to repackage data into a workable format.

⁷⁴ Interface in the technical sense, but also the semantic sense, a point which is of central importance throughout this paper.

⁷⁵ Pages 5 - 7.

- ***What about statistical uses?***

The shift from the administrative world to the statistical world is one example among many of the challenges of transition, of moving data between different specialised domains: the need to package data, the problem of *data friction*, and the need to detach data from their original ecosystem and reattach them to a new one.

There are more specific issues to be dealt with, too, one upstream of the transition on the supplier's side, and one downstream on the user's side.

On the supplier's side, it is important to remember that the concept of *data friction* originated in the scientific domain (Edwards et al. 2011), where we can assume, simplifying matters slightly for the sake of argument, that we are dealing with frozen data, as they represent the recorded results of experiments. Administrative data, however, which are liable to evolve regularly in response to events (births, marriages and deaths, user requests, administrative decisions etc.), are often anything but frozen. Detaching these data from their original context thus requires an operation to freeze the data, along with all the other work involved.

On the user's side, statistics are created and used for information purposes, structured by the analytical framework set out in our first chapter (units, variables, scope, time, categories): while questions about units and variables will inevitably arise whatever the proposed use of the data, the extent to which the scope is complete will play a major role, as will the quality of classification. Special attention will also be required for data with dual time frames. Last but not least, statistical uses allow for the existence of a "margin of error" which is permissible, and even measurable, unlike the statistics generated for accounting purposes (where it must be possible to verify every last penny), or other uses with operational implications (the value of your pension, for example).

In all cases, the key is to ensure that all parties speak the same language, abide by shared conventions, and remain attentive to any implicit assumptions.

3.e. Statistics based on administrative data: a three-phase process

Since statistical analysis is not the original purpose of these data, they must be transferred out of their native information systems. The very foundations of statistical estimates thus disappear, and a set of transformations is required to make meaningful estimates possible. A detailed assessment of the quality of administrative data is also indispensable: Statistics New Zealand (2016) offers a very comprehensive guide to this process, analysing potential sources of error.

In order to transform "initial" administrative data into data that is equivalent to the results of a survey, we propose a three-step phase process in which each phase is devoted primarily to resolving one or more of the main difficulties encountered.

The first phase, which we call **acquisition**, seeks to constitute a "true" administrative source of reference. This phase incorporates the definition and realisation of the freezing operation discussed above. It concludes with a quality assessment exercise which constitutes a "go/no go" for the rest of the process. By the end of Phase 1 we should find ourselves with data which are comprehensible to the administrative entity, and which allow us to conduct verifications and ask questions. It is important to note that while designing this acquisition phase, we must always remember the subsequent uses we have in mind for the data, i.e. the type of statistics we hope to produce. The acquisition phase is the subject of Chapter 4.

In the second phase, which we call the **transformation** phase, we intentionally detach ourselves from the administrative realm in order to reattach ourselves (as Courmont might say) to the world of statistics. We thus find ourselves in the other branch of the Y diagram (see Section 2.e), doing work that is specific to administrative data, something which is far from self-evident: as we have seen (Borgman, Groth 2025), one does not simply flit carefree from data creator to data user, from one domain to another.

The transformation phase will be the subject of Chapter 5 and naturally concludes with a data quality assessment process, albeit a process which in no way resembles Phase 1: this time we are operating firmly within the realm of statistics, with its objects, variables and scope, all of which must be readily comprehensible to statisticians. The end of Phase 2 corresponds to the fork in the “Y”, after which the processing operations will be the same as those used in the context of surveys. Positioning ourselves in relation to the GSBPM, these two phases of acquisition and transformation run parallel to the design, construction and collection phases applicable to surveys.

They enable us to pass from one world to another, a movement summarised in the table below:

What exists in the administrative sphere		What we need in the statistical world
Live information system		Frozen dataset
Very rich, disparate information system		Observations and information which is useful for statistical purposes
No collection period		
No reference period	→	Reference period associated with all data
Lists of administrative codes		Classifications
Administrative objects		Statistical unit(s)
Administrative concepts		Variables of interest
		Belonging to statistical scope
Real administrative scope		Representativeness of statistical scope

Having now entered the statistical world, we can move on to Phase 3, which we call **statistical processing**, and which closely resembles the work done during surveys (in fact, this is precisely the processing phase found in the GSBPM). There are a number of implications for administrative data (including some functions found in surveys which now disappear) which merit further explanation, and further details can be found in Chapter 6.

B. Our proposed approach

4. The acquisition phase: evaluating the quality of administrative sources

4.a. Principles

During this first phase, we must resolve several of the defects inherent to administrative information systems, in order to facilitate future statistical use of the data.⁷⁶ We must thus deal with the following issues:

- The live character of the source IS;
- The great size, and complexity, of the source IS;
- The lack of clarity in the temporal dimensions (collection period, reference period);
- The heterogeneity of the files retrieved.

Acquisition thus involves freezing the administrative data, using our knowledge of the reference date/period associated with the information to perform operations such as: filtering for objects or attributes (we do not need *all* of the available data), connecting multiple spreadsheets from the source IS, merging files (for example, files from multiple counties collated into a single document), but *not* transforming the data.

The source which emerges from this first phase is still aligned with the concepts, objects and domains of the administrative agency. This is a defining characteristic of Phase 1, ensuring that statisticians and representatives of the administrative entity speak the same language, and are able to exchange in a manner which will no longer be possible once the data have undergone transformation.

The acquisition phase is by no means limited to “retrieving a frozen, clearly defined source.” We also want to gain a clear understanding of its quality. The acquisition phase thus involves a whole process of quality assessment, including much back-and-forth with the administrative entity. At the end of this phase, we find ourselves with a clearly defined and evaluated administrative source, a far cry from the nebulous IS we started with.

Our proposed approach to achieving this transition can be split into two main phases, drawing inspiration from Cerroni, Bella & Galie (2014) and Daas (2009). These authors argue for a clear distinction between analysing the source and the metadata, on the one hand, and the work performed on the data themselves, on the other hand (cf. **Annex 2** for how this is applied at Istat).

The first phase is **preparation**. It is all about discovering the administrative context: what it means, how it functions, what its legal framework looks like. These are not mere abstractions; they are inextricably bound up with the domain knowledge, its objectives, language and practices, and it is important to grasp these codes. All of which requires access to detailed documentation and a good understanding of the metadata (regarding the value domain, for example), which will be indispensable for the next steps.⁷⁷ The preparation phase also includes a concerted effort to establish a pipeline for data transmission: protocol, interface contract, data contract etc.

Data receipt and quality assessment make up the second phase. This time we retrieve the actual data, while conducting a series of automatic (or manual) checks. To do so, we must have established in advance exactly what we mean by “quality”.

⁷⁶ It is important to note that we are talking here about individual data, but there may be certain cases in which data arrive in aggregated form. This possibility is not explored in the present working paper.

⁷⁷ See also Bonnans (2019) on the subject of the metadata used in France’s official statistics system.

Splitting the process into two phases does not mean that it is necessarily a linear progression: during the quality-assessment phase we may discover problems of various kinds (invalid codes, inconsistencies, implausible value distributions etc.), which will need to be discussed with the administrative entity. This is where the idea of “feedback loops” comes into play, allowing us to improve the resulting files. These files are both an *output* of the acquisition phase and an *input* to the transformation phase.

Last but not least, the process described above, with its two phases and feedback loops, will naturally look very different if we use the same source multiple times: this is true when the same statistical producer makes repeated use of the same source, and also when usage is pooled among multiple producers. We thus consider the consequences for the acquisition process.

4.b. The preparation phase: source, ecosystem, contracts

The existing literature on administrative data highlights the importance of adopting a comprehensive approach, encompassing not only the data themselves, but also the process by which they were produced and the organisation responsible for that process, who we will call “the supplier”.

- **Suppliers**

Suppliers are defined by their status, their position, and the legal framework within which they operate.

Naturally, the public finance directorate is responsible for fiscal data, which are the core of its activities. These data may also be held by an operator, who is not professionally responsible for them, but manages the necessary data flows: this is true of the GIP-MDS⁷⁸, responsible for a vast set of social declarations, including the DSN, but also employment declarations (DPAE⁷⁹) and MSA⁸⁰, URSSAF⁸¹ and PASRAU declarations.

When examining a potential data source, and once we have determined which organisation could legitimately fulfil the role of “supplier”, we must decide whether or not the organisation in question is reliable *in this role*. For administrative declarations, for example, this reliability is derived from the existence of legal obligations incumbent upon declaring parties.

This is true of individual social declarations (DSN), but also the FINESS register of healthcare establishments, where all modifications must have a solid legal basis (Bensoussan, Bizingre, Courvalin 2023).

We must also consider the long-term viability of the source, which is not a foregone conclusion (cf. the abolition of council tax in France), and even the long-term future of the supplier.

78 French Public Interest Group (GIP) ‘Modernisation of Social Declarations’. Created in 2000 to meet the need to simplify administrative procedures for companies, GIP brings together all the players in the social protection sector.

79 Déclaration Préalable à l’Embauche - Declaration prior to hiring

80 MSA (Mutualité Sociale Agricole) provides social security coverage for all farmers and their dependents in France.

81 The “Union de recouvrement des cotisations de sécurité sociale et d’allocations familiales” (Urssaf) is a French organization. Its main missions are to collect employee and employer social security contributions to finance social protection (pension funds, unemployment insurance, etc.)

- *A potential source: the ecosystem*

Of course, you cannot just “supply” data in the same way that you might “supply” tables and chairs, for example. As we saw in Chapter 3, before obtaining a file we must take a deep dive into the universe in which the data were created, considering the reason for their existence, and studying the problems encountered by the administrative body whence they originated.

All of which requires a certain degree of curiosity and open-mindedness from statistical institutions. These attributes are essential to understanding how data are managed, which objects are involved, how they relate to one another, the concepts and history which informed their creation etc. These discussions take time (cf. *data friction*), and may even require the secondment of “embedded” statisticians to work within the administrative bodies supplying the data.

Secondment is an intensive solution employed by the UK’s Office for National Statistics (ONS), with statisticians embedded with suppliers for several months. The ONS can thus get involved in the co-construction of the source produced at the conclusion of Phase 1 (Fermor-Dunman, Parsons 2022).

Understanding the data also requires some knowledge of how they were put together. Zhang (2012) proposes that we should seek to identify the different types of error⁸² which may arise when data are created, or later on in their life cycle (Shah et al 2021):⁸³ the aim is to focus on those points where the administrative process could diverge from the theoretical objective. This analysis is invaluable, helping us to correctly identify the criteria which can be utilised to optimise the data.

The need to develop our understanding of the environment and the source remains a constant priority. This is true throughout the duration of the preparation phase and beyond. Nonetheless, caution is required: this is not an exercise to be conducted for its own sake; the aim must always be to **respond to a clearly defined statistical need**. This brings us back, from a slightly different angle, to the **definition of needs**, the first phase in the GSBPM which is inextricably connected with this exploratory analysis of the source. In practice, either a need exists and we wish to ensure that the source contains relevant data to satisfy that need, at least partially, or else the data themselves open up new opportunities.⁸⁴

Whatever the case may be, a close analysis of the administrative objects and associated concepts, along with an understanding (at the very least a superficial understanding) of the specific processes employed will prove invaluable when it comes to determining the structure and content of the frozen data file, particularly when no such file yet exists within the administrative body providing the data. We should also note that there are some situations, and fortunately so, in which statisticians are thoroughly well versed in administrative data, their environment and their specific uses. This is particularly true of sources which have been used for statistical purposes for a long time (but nothing should be taken for granted, as sources may change over time).

82 See page 43. The author is clearly putting himself in a statistician’s shoes. But the term “error” merits further parsing, as not all observers define it in the same way.

83 Some caution is required with regard to the “data life cycle”, a notion for which there is no standard definition, nor even a broad consensus. The 2021 article cited here lists 76 known models...

84 In the private domain, examples include the use of mobile telephone data to estimate the number of people leaving Paris during the Covid lockdowns. For administrative data, examples include the use of the DSN to measure partial unemployment:

see <https://blog.insee.fr/suivre-la-conjoncture-lorsque-les-entreprises-repondent-moins-aux-enquetes/> in french.

- ***Retrieving metadata about sources***

Now that we have a firm grip on the ins and outs of our source material, we can move on to an analysis of an entirely different, and more practical, nature: seeking out all of the information, some of which will be technical, required to better understand the contents of the source, and how it was put together. We need data about the data, which we call metadata. In this hunt for documentation, the nuggets we unearth will be relevant to our analytical framework: definition of objects, variables, domains of values, scope and time frames.

The example of the DSN, with its “technical specifications,”⁸⁵ offers a comprehensive and rigorous example. The scope and time frame (declaration dates and reference dates) are described in the introductory pages of this document. The concepts employed are derived from a conceptual data model, which yields declaration “blocks”. We specify the variables, or “headings”, which are attributes of these concepts. For each variable, the value domain is defined by different characteristics: classification, list of values, inclusion in external registers, and even the use of *regular expressions* formally defining the structure of a character chain (Friedl 2006). All of these elements add up to constitute an “exchange standard”, with annual versions shared among the various partners and backed with dedicated tools (Dubrulle, Rosec, Sureau 2023).⁸⁶ Allow us simply to note that, strictly speaking, this documentation pertains to the data flow, not the source which may be built upon these foundations. Moreover, the formal description is important but not sufficient to protect against quality issues related to declaration practices (declaration of paid hours, for example).

The chronological separation between understanding the source and obtaining the metadata is somewhat artificial: often, work on the metadata begins during the exploratory analysis of the source, and continues through interactions with the data supplier. Exploratory analysis may thus draw upon the metadata, where they exist, and benefits particularly from productive dialogue with suppliers (Eurostat, 2017b).

Drawing parallels with the GSBPM, this phase has much in common with the **design step**.

- ***Establishing the practical means of data exchange***

Ultimately, the goal of the preparation phase is to organise the transmission of a “source” from the administrative agency which constitutes the “supplier” to the statistical office which acts as the “recipient”... with a view to becoming a producer of statistics.

With this goal in mind, there will be numerous exchanges between data supplier and recipient, involving technical directors as well as IT specialists. Together they must establish secure and robust processes for data exchange, as the files are often very large and must remain confidential.⁸⁷

We must thus define the terms of **cooperation** between the statisticians and the public body (Chaleix, Vanderschelden 2023), an arrangement whose relational dimension should not be neglected. The relationship may take many, non-mutually exclusive forms, from email exchanges and occasional meetings to more formal negotiation processes. To regard administrative agencies simply as “data providers”, would be to overlook one key thing: we have not really defined what is

⁸⁵ <https://www.net-entreprises.fr/media/documentation/dsn-cahier-technique-2024.1.pdf>

⁸⁶ Another noteworthy feature of such agreements is that the standard is updated each year in accordance with a precisely defined calendar, giving official statisticians plenty of notice and time to adapt their information systems to any changes.

⁸⁷ This does not always go smoothly, and even in 2025 it is not unheard of for paper copies to be sent as a back-up solution...

to be provided, and what that actually means, and therein lies the real problem. Such discussions may be easier if you are in direct contact with the statistics team at the administrative agency.⁸⁸

In France, some public bodies have their own statistical offices: they include various national social security agencies, such as the CNAF, CNAM and CNAV.⁸⁹ These in-house statisticians often act as intermediaries in the exchanges between producers of statistics and the administrative entities who supply the data: allowing statisticians to “talk amongst themselves” allows for a greater degree of understanding, reducing the risk of *data friction*.

The question of the legal framework applicable to data transmission will also need to be addressed. In France, thanks to Article 7bis of the law passed in 1951⁹⁰, we have such a framework for administrative data, but it has its limitations. The issue of how to manage confidentiality will also need to be resolved (Redor 2023).

As these and other issues are addressed, a **contractual** relationship gradually begins to take shape between the administrative entity which holds or supplies the data and the organisation(s) planning to use those data. The process culminates in a written agreement binding the parties. This text provides a clear definition of what is to be transmitted, by whom, when, how, by what means and in what format.

One good example is the agreement put in place between INSEE and France’s Public Finance Directorate for the use of different tax datasets (with an annex for each source).

In some respects, the contract negotiation process provides a way for official statisticians to take back a measure of control over administrative data. To that end, the existence of formal agreements is indispensable: they must be very precise, particularly with regard to delivery dates (deadlines to be respected!) when the source is an input to a process of statistical production which has its own constraints.

Example: the DSN is used in the production of economic outlook data, whose publication calendar is stipulated by Eurostat. Late delivery of the data may lead to a delay in publication, and thus an infringement of the agency’s obligations to Eurostat.

In cases where data are used on a one-off basis, there may not actually be any formal transmission process. In fact, there may be no written agreement whatsoever: in these circumstances, statisticians (or data scientists) strive to “do their best” with datasets which they discover as they go along. This is a common state of affairs in the “data departments” of government inspectorates, whose missions are time-limited: once the work at hand is complete, the majority of the data are no longer of any interest (Bolliet et al 2025, Berthe 2025).

- **Results of the preparation phase**

In terms of concrete outputs, the preparation phase leaves us with:

- A set of metadata,
- Documentation,
- A form of contractual agreement,
- ... and a decision to be made on whether or not to continue with the process.

88 Statistical offices may thus end up as clients of their own administrative entities, obliged to deal directly with the difficulties inherent to freezing information.

89 CNAF = *Caisse nationale des allocations familiales*, or National Family Allocations Fund; CNAV = *Caisse nationale d'assurance vieillesse*, or National Old Age Insurance Fund; CNAM = *Caisse nationale d'assurance maladie*, or National Health Insurance Fund.

90 Article 7 bis of Act [51-711 of 7 June 1951 on statistical obligations, coordination and confidentiality](#)

One operational choice open to statisticians at the conclusion of this evaluation process is to decide that the project is not worth pursuing any further, because the source does not meet their needs. There may be any number of reasons for this: there may be statistical considerations in play (concepts or objects which differ too much from the units in which we are interested, a source whose theoretical scope is too limited, data which take too long to access), not to mention potential ethical reasons (disproportionate privacy issues when using personal data, or the acceptability of the project to civil society).

Examples include the Résil programme, which decided not to make use of health data (use of health insurance cards), RNCPS⁹¹ data (a register used to combat fraud) or data derived from residency permits, in order to improve the social acceptability of the study, sacrificing some of the statistical quality which had been demonstrated by experiments with data samples (Dupont et al., 2024).

If we do decide to pursue the project further, the preparation phase will have laid the groundwork for the next step, data receipt and quality assessment. The purpose of this next phase is to obtain, for the subsequent phases, data whose level of quality is known to us.

Before we delve further into the source evaluation process, it is worth asking a simple question: what exactly do we mean by “quality”?

4.c. The quality of administrative data: an unstable notion

- *The notion of data quality*

Quality is not an absolute truth: it is measured **with reference to a given purpose**: “*Quality is fitness for use.*” (Juran 1951) Take the example of cars: our definition of what constitutes a quality automobile will differ depending on when we are heading off on a family holiday, or else attempting to break speed records on a racetrack.

The same is true of data quality: “*Quality is not a property of the dataset itself, but of the interaction between the dataset and the use to which it is put.*” (Hand 2018). Even if some markers of data quality are generally applicable (e.g. compliance with format), there is no “inherent” notion of data quality. This may lead us into some paradoxical situations, with data which are regarded as fit for one use, but not for another. Edwards (2010) illustrates this difficulty in his description of mutual misunderstanding between meteorologists and climatologists, who use data for very different purposes, and thus have very different quality criteria.

A second important idea: data quality is multidimensional. In the academic literature on this subject, the components of quality which crop up are always more or less the same.

Di Ruocco et al. (2012) identify different families of indicators: relevance, accuracy, completeness, consistency, temporal accuracy, accessibility, interpretability, uniqueness, coherence and compliance with a standard. Bontems and Goulin (2013) limit themselves to 6 dimensions: pertinence, accuracy, the speed with which information is disseminated,⁹² accessibility, ease of interpretation, and coherency. Batini and Scannapieco (2016) introduce the notion of database *schema* quality, insisting upon the importance of reaching a compromise between the different

91 Répertoire National Commun de la Protection Sociale (Common National Social Security Register)

92 From the statistician’s perspective, this speed corresponds to the gap between the reference period and the data acquisition period. For users, the quality of statistics is partly dependent on their timeliness, and is thus dependent on the publication timeframe, downstream of the acquisition phase.

dimensions. Loshin (2011) distinguishes between intrinsic dimensions (accuracy, *lineage* or traceability, coherence and homogeneity of formats, semantic coherence – existence of definitions, uniqueness of naming system) and contextual dimensions (completeness, coherence, timeliness and punctuality...). Similar notions are deployed to define the quality of *statistical data* in the code of best practice published by the EU's statistical agency (Eurostat 2017a).

- ***What about the quality of administrative data?***

Like any other data, administrative data are designed for specific uses. These uses have their own quality criteria, which differ from those prioritised by official statisticians. Transferring administrative data into the statistical realm thus requires us to verify their suitability for new uses.

This is about verifying the quality of the intrinsic dimensions of the data (since the acquisition process was conducted by a third party), but also the contextual dimensions because the data are now “detached” from their original use, and set to be reused for statistical purposes.

In the United Kingdom, the UK Statistics Authority has defined the central pillars of *Quality Assurance of Administrative Data* (QAAD), including the operational context in which data are retrieved from administrative sources, as well as communication with data suppliers and engagement with their quality assurance principles (Babb 2017). More recently, a general quality assurance framework was adopted by the UN Statistical Commission, comprising 10 critical requirements (United Nations 2025).

The fact that administrative data are used to support administrative activities does not guarantee that they are of sufficient quality for statistical purposes. They may not have been updated for some time, for example because the interested parties failed to submit the necessary information.

This is true, for example, of the “Address” variable in the administrative data provided by the Northern Ireland health authority: having analysed a vast sample, Foley et al. (2018) conclude that certain sub-populations are less likely than others to inform the relevant administrative authorities when they change address (young people, for example), leading to numerous errors.

Data may also show traces of **bias**, another error risk induced by the fact that it is sometimes “beneficial” to declare a certain value to the administrative authorities: in order to obtain a financial allowance or subsidy, for example. This is not a matter of intentional deceit, but rather of opting for the most favourable interpretation available.

Example: in certain circumstances, the ‘primary activity’ declared by a company may have operational implications, including determining eligibility for some subsidies.

At a more general level, and contrary to the way things work in statistics or with scientific experiments, administrative entities do not regard data as an end in and of itself. As we saw in Chapter 2, more often than not data merely exist to serve as action triggers, go-betweens or by-products. As such, from an administrative perspective, there are no quality problems as long as the data allow for the right action to be triggered at the right time.

Example: for the pension funds, a “career” is a trigger variable for the action “calculation of pension rights”. If details of this career are incomplete, but if the gaps in the information change nothing in terms of pension entitlement, then the career data are “of good quality” for administrative purposes. Researchers, however, will take a very different view. They need access to exhaustive information in order to conduct longitudinal analyses.

As a result, it is entirely possible for data to be regarded as being of sufficient quality for administrative uses, but for those same data to be dismissed as being of poor quality for statistical purposes.

In other words, “quality” is not synonymous with a low or non-existent error rate: where an administrative agency may not detect the slightest error or dysfunction, statisticians may find administrative data to be riddled with all sorts of defects.

- *A few examples of divergence between statistical and administrative definitions of quality*

The following examples pertain to different dimensions of quality.

Approximate information

Administrative bodies do not always need their data to be precise, particularly when it comes to data from the more or less distant past.

Example: in the land register, the “year of construction” data for the 19th century shows distribution peaks in years ending in 0. This seems to be due to the way in which information was recorded in this era. This does not represent a quality problem for the land registry department, since it has no direct operational impact. However, for anybody wishing to conduct a more detailed temporal analysis, for example with a view to gauging the impact of historical events in this century (1830, 1848, 1870 etc.), this level of quality would be considered wholly insufficient.

Information in formats unsuitable for statistical use

The information contained in sources may be accurate, but presented in a format which makes it difficult to use, particularly when dealing with large volumes of data.

Examples: in the vehicle registration files, information regarding the names of owners is contained in a single field, which leads to processing difficulties in cases where vehicles have multiple owners.

Sometimes, information is contained in PDF scans of paper documents, obliging us to use optical character recognition (OCR) tools to decipher the text in the images and obtain the data. Denis (2018) provides an example involving data on the location of cycle lanes within cities.

Data conventions open to misinterpretation

In terms of assigning values to data, administrative entities sometimes have their own conventions which, seen from the outside, may seem strange.

Example: in the land register, the field indicating which floor a premises occupies can be somewhat surprising. Analysing the distribution of this variable reveals a peak... for premises on the 81st floor!⁹³ Are there really so many buildings of 80+ floors that have previously gone unnoticed? No: the code 81 is simply used as shorthand for the first basement level, or Floor -1. Of course, database designers are fully entitled to make use of such shortcuts. We can easily imagine that the field was only capable of holding positive

⁹³ As we shall see later on, it is always instructive to look at the *effective* distribution of variables within the files we retrieve.

integers, and so, to avoid rewriting the code, they decided to use a large (and very rare) value to replace the negative sign and denote subterranean levels.

Missing updates

There are also situations in which no errors were made when recording the information, but the information in question has not been updated since. If you think about it, this is entirely natural.

Example: an establishment changes its telephone number or contact email address. There is no way of ensuring that this information is updated in the IS: we cannot expect administrative agencies to constantly scrutinise and check all establishments by means of daily questionnaires.⁹⁴

Inevitably, this temporal issue leads to inconsistencies in the data. Blame may be laid at the door of the administrative body (who cannot check everything all of the time), but also of users.

A common example: people who submit doctor's notes extremely late. Until the note has been received, as far as the administration is concerned no medical appointment ever took place, because it does not exist in the information system.

Problems of this kind are frequently encountered when recording those entries which are removed from data registers. Updates are often asymmetrical: new entries are always recorded, especially when registration is essential to access certain rights (e.g. the right to do business), whereas failing to report the fact that they have exited the list does not have adverse consequences for the interested party.

Default values, or how to deal with incomplete information

Administrative entities may find themselves in a situation where the information at their disposal is incomplete, inaccurate or simply lacking. They may respond to these challenges in different ways, some more formal than others.

Example: an administrative body does not know the date of birth of a particular individual. In practice, they may choose to record values which are outside the acceptable scope for this variable: 00/00, 99/99, or WW/WW. These act as codes indicating the absence (or inaccuracy) of available information, but the codes are often not standardised, and may vary from one source to the next. It is also possible to use values from within the acceptable scope, but this makes it even more difficult to identify incomplete entries. For example, when the date of birth is unknown a clerk may enter 1st January (01/01) by default. In this case, it is impossible to tell the default responses from the "real" 1st January birthdays.

4.d. The data receipt and quality assessment phase

Following on from the preparation phase – which enabled us to better understand our data source, making use of documentation and metadata – it is now time to get to the heart of the matter: the data themselves. Cerroni, Bella & Galie (2014) highlight the importance of "evaluating" administrative data once they have been transplanted from their native environment, and before embarking upon

⁹⁴ Unless such changes require a declaration to the relevant administrative body.

statistical transformation.⁹⁵ As we have seen, however, gauging the quality of administrative data is no easy matter.

We thus have a “provider”, an administrative entity all set to transmit datasets to the “receiving” statistical office. In technical terms, these datasets will take the form of one or several files. In some cases, the datasets will have been constructed entirely by the administrative entity; in other cases, they will be the fruit of collaboration with the statistical service. In practice, the transmission of data by an administrative entity is never as simple as sending a single file: each *instalment* may comprise multiple files, and delivery of the data in their entirety may be spread across multiple instalments.⁹⁶

Example: the DSN and PASRAU administrative declarations are transmitted in several instalments: twice monthly for the former,⁹⁷ daily for the latter. The period of reference for both declarations is monthly.

During the reception – assessment phase, we seek to verify the technical solidity of Phase 1 as a whole, from compilation of the dataset through to transmission, in order to prepare effectively for the subsequent phases. To do this, we make use of quantitative measurements to evaluate the quality of the data. Ideally, these measurements should be defined in advance, based upon a solid understanding of the administrative process and the stated statistical needs.

- ***Compliance of deliveries***

We want to be sure that the files transferred are, on the whole, consistent with our expectations, especially in quantitative terms.⁹⁸ There are various ways of checking deliveries, particularly by making use of indicators which allow us to gauge the volume of the data:

- number of files shared (compliance with the terms of the data sharing agreement, or a more detailed list of the files we expect to receive),
- number of entries: this provides some indication as to whether the delivery complies with expectations, and also allows us to check that the volume of data is of the expected order of magnitude for the relevant *scope*.

For example, if we are expecting to receive data relating to 1.2 million declaring parties, we must check that, once all of the deliveries have been made, we are close to that number.

In practice, administrative sources are usually sent as multiple files: they may be files corresponding to distinct administrative units (e.g. a department within an administrative entity, or a regional agency), or files corresponding to different periods. The number of files expected, as specified by the administrative body, is an important piece of information: we must check that it matches up with the number of files actually received.

For example, if an administrative body were to promise us a separate file for each department in France, then even an instalment containing 50 files would be incomplete⁹⁹. In some cases, administrative bodies directly specify the number of files they plan to send, or, if they wish to be much more precise, they may provide a list of names.

95 N.B. this does not mean that the data have not been transformed, that they are somehow “raw”. As discussed above, and by Gitelman (2013) and Denis & Goëta (2014), among others, the term itself is an oxymoron. In all cases the data received will have been transformed, in one way or another, *within the administrative entity*.

96 The technical specifications of the DSN include a definition of what constitutes an instalment.

97 From 2025 onwards.

98 An all too familiar problem when defining a data exchange protocol.

99 There are 100 départements in France.

The advantage of these initial, very basic, verifications is that they can be implemented rapidly. This saves precious time: there is no point launching the whole costly process of data uploading if, for example, only a third of the data has been transmitted correctly. The aim is to be responsive, contacting the supplier immediately if a problem is detected with the data transmission (files corrupted during the delivery process) or the preparation of the files by the supplier (incomplete files etc.), so that the latter can rapidly remedy the situation with a follow-up delivery, if necessary. Responsiveness is of the essence in cases where administrative data constitute an input to a statistical forecasting process which runs to a very tight deadline (as is the case with the DSN).

When checking the compliance of a delivery, we must also check some *paradata*: these are auxiliary data pertaining to the process of data transmission, reception and downloading, which may be produced by the departments responsible for receiving the data.

- ***Content compliance: domain***

For each delivery, we conduct additional automated controls allowing us to:

- Verify the format: file structure and classification (alphanumeric code, text, integers, dates etc.),
- Verify, for each *variable* considered to be of interest for statistical purposes,¹⁰⁰ that the *domain* indicated in the data sharing agreement / documentation is respected: data within an accepted range of values,¹⁰¹ intervals, respect for a specific structure in the character strings etc.¹⁰²

All of these checks are conducted automatically, as administrative sources generally contain far greater volumes of data than we are used to working with when conducting surveys – sometimes several million individuals/objects, rather than several thousand. For this reason, automation is indispensable.

By way of an example, the Résil programme has put in place a dedicated system for managing the reception and availability of demographic and social data, using a generic tool for receiving and checking administrative data, known as ARC (Lefebvre, Soulier, Tortosa 2023) and capable of conducting automated operations.

When working with such large volumes, we do not have the luxury of *manually editing* the microdata as we do with business surveys, where there is a precisely calibrated combination of automatic and manual checks (Hedlin 2003).

If these compliance checks are not met, as per the conditions stipulated in the data sharing agreement, the data delivery will not load properly. In such cases, the statistical office may deem it necessary to seek corrections or explanations from the providing party, or even request a second delivery.

- ***Plausibility of distribution***

We now move on to a subtler form of quality assessment, with an exercise which cannot be fully automated and requires informed interpretation. By studying the distribution of the values for the

100 Assuming that they have been defined. We do not set out to check all of the administrative data, which would be a waste of time, focusing instead on those variables likely to be of interest for statistical purposes.

101 The term used to indicate “data missing” (for example, 9999) is considered to be within the scope.

102 For example, registration numbers and dates are series of characters, arranged in a specific order. See Dubrulle et al. (2023) for examples, or Friedl (2006) for an explanation of “regular expressions”.

different variables, we can ensure that nothing abnormal is going on. Once again, the object of our enquiries is the domain. In the examples given below, we shall limit our focus to cases where the domain is a predefined list of values.

Missing values

This first example of a potential error does not require any subject-matter knowledge. Imagine that the domain of possibilities for a variable X is {a, b, c, d}, and that in the file received we find only 'a's and 'd's. Values 'b' and 'c' are missing entirely, which is surprising, even if it does not prove incontrovertibly that an error has been made.

Example: the first batch of data received from the GMBI source contained exclusively properties for which the variable "Nature of occupation" showed the value "Premises occupied". The absence of any other values revealed that an error had been made when selecting the data.

Distribution peaks

We may find that certain values within the domain of possibilities are not absent, but in fact observed so frequently as to be excessive.

Date of birth, already mentioned above, is an obvious example: certain dates appear more frequently than we would expect, such as 01/01 or 31/12. We also often find that births peak in the year 00. This probably suggests that the administrative body in question does not actually know the correct date. This is not really an error, but rather a useful piece of information for later use.

Credibility of distributions

A more difficult task, and one which is even harder to automate, is to assess the credibility of the distribution of "interesting" variables. When acquiring data for the first time, we do not always have pertinent points of reference allowing for robust *a priori* comparison of the distribution of these variables. Nevertheless, if we have data for all of France's *départements*, for example, we might reasonably expect the age and gender distributions for each *département* to be relatively close to those of the general population, if the source covers the entire population. It is therefore always useful to check the distributions of known variables, to ensure that they are not abnormal.

As we shall see, distribution may also be checked much later on in the process, in Phase 3, so as to avoid the need for corrective measures.

- ***Scope: overcoverage and undercoverage***

While it is hard to be sure that the intended scope (of a source, as defined by the administrative entity) is amply covered by the file(s) provided, we can at least check that there are no obvious problems.

We can start by performing some simple volume checks (counting the number of objects), looking closely at developments from one period of time to the next: if the source is not new, and we already have data covering earlier periods, then we can compare aggregates from different deliveries and identify any problems or changes in the data.

Example: when transferring to the new GMBI source, the tax authorities decided to assign individual registration numbers to all subsidiary buildings which were previously considered as part of the main premises to which they were attached. The number of registered business premises thus exploded from one year to the next. This increase was in fact “normal”, as it simply reflected a change in the data management. Nonetheless, a new filter had to be added in order to pick out the statistical target population.

We may come to realise that a section of the expected scope is missing from the file, simply by observing that certain objects are missing, or present in unusually small quantities. We call this *undercoverage* of the scope. In practice, undercoverage is always difficult to spot because “we don’t know what we don’t know.”

Example: in some instalments of data on key facilities (Helfenstein 2022), there have been certain French counties without a single recorded petrol station – which seems unlikely.

At the other end of the scale, we sometimes find ourselves with files containing units which fall outside the theoretical scope of the administration: this is *overcoverage*. In much the same way described above, such coverage errors can be detected thanks to the presence of unexpected units.

Duplicate entries represent a very specific form of overcoverage: the scope is exceeded because the same unit has been counted twice. For example, in an information system where a given operation triggers a given action, we may come across cases where the same operations (and ensuing actions) have been performed twice. This may not be of any importance to the original information system, but it nonetheless creates duplicate entries which cause problems for subsequent statistical uses.

- ***Quality-assessment tools***

All of the processing operations performed on data (checking formats, inter-variable checks, distribution checks and the like) can be grouped under the general heading *data profiling*, thoroughly described by Olson (2003). Data profiling techniques, but also tasks such as standardisation and double entry removal, are deployed with the help of *data quality tools* (Van Dromme et al. 2007, Boydens 2021), available on the market or in open-source mode. As Boydens, Hamiti & Van Eeckhout (2021) explain, these tools are synonymous with a curative approach to data quality.¹⁰³

Nevertheless, when seeking to determine the true quality of data files received from administrative sources, statisticians find themselves working “without a safety net” because they have little to which they can compare their findings. Intuitively, we might regard true quality as a combination of compliant values, plausible distribution (for each variable), macro-level plausibility (rate of missing declarations, missing values) and the absence (or low rate) of duplicate entries, and perhaps even the likelihood of the combined distribution curves.

Comparisons between data are common enough in the statistical world, but at this stage in the process we are still very much operating within the administrative realm. Evaluating quality with the help of individual or aggregated data from another source is something we will discuss in the next phase.

¹⁰³ This is entirely complementary with the *preventive* approach to quality described by Boydens et al. (2021) in the article cited: this is a very interesting approach from the administrative perspective, but falls beyond the remit of the present paper.

- **Comments**

When evaluating data quality, we must also be wary of things which can be improved indefinitely, and thus avoid the fate of the Danaids trying (and eternally failing) to fill up their leaking jug. Hence the importance of clearly stating our statistical needs from the outset, i.e. our expectations in terms of the quality we wish to see in the resulting statistics. What level of detail do we hope to achieve? What are the expectations of our users in terms of the timeliness of the data? Which statistical units and variables of interest will we be including?

As such, and this may appear counter-intuitive, evaluating the quality of a dataset is not about identifying every last anomaly, improbability and error. It is, in fact, a matter of ensuring that the data reach a level of quality *consistent with the objectives of the statistical products* we intend to produce. Returning to the example of construction dates mentioned above (many buildings are recorded as having been constructed in a year ending in 0), there are many uses where this lack of accuracy is of no importance whatsoever.

Only once we have defined these expectations can we break them down, as far as possible, into measurable indicators and quality targets, providing us with a framework which will be particularly useful when we come to use it again. This framework can be further broken down into tools.

4.e. Feedback loops with administrative entities

Since administrative data always contain defects, in one way or another, the quality assessment will generally reveal a number of anomalies. Contrary to survey data, administrative data do not give us the option of checking seemingly anomalous results with the unit in question (individual or business): the only feedback at our disposal is that of the administrative agency providing the data, and they will not be able to provide answers for individual entries: their feedback will operate at a more *macro level, pertaining to the file or the delivery as a whole*. This has some fundamental consequences for our verification efforts.

We use the term “feedback” loop to describe the whole process of back-and-forth between administrative entities and statistical agencies, which may culminate in new deliveries (of data or metadata) or improvements in the quality of administrative sources for future instalments (Renne, 2018).

- ***A precondition: evaluate the quality of administrative data in their own context***

In the preceding section, quality assessment was shown to focus on objects and variables as defined by and known to the data provider. This placed us firmly in the administrative realm, rather than the statistical realm, and this is no minor detail. Once data have been transformed for statistical purposes (as we shall see in Phase 2), they become illegible to the public body whence they originated, ruling out any further discussion.¹⁰⁴ Adopting the vocabulary of the administrative world (*taxpayer* rather than individual, *tax household*¹⁰⁵ rather than household etc.) may seem restrictive or even artificial for statisticians, but experience has shown that it is a necessary condition for constructive collaboration.¹⁰⁶ This is a key point.

Statisticians must adapt to the administrative context, and not vice versa.

¹⁰⁴ Also known as semantic continuity between data providers and recipients, closely related to the concept of *data friction* (when discontinuities arise). A related concept is the notion of distance between producer and user, mentioned above in relation to Borgman & Groth (2025).

¹⁰⁵ A tax household (foyer fiscal) designates the individuals who are grouped together on a single income tax return.

¹⁰⁶ This is not purely a matter of terminology; there is also a semantic dimension to the language employed.

This approach allows us to:

- Better identify the origins of problems in the data (changes in data quality, transmission issues, processing problems etc.);
- Converse with data providers using shared terminology, concepts, objects and variables. Such discussions are essential to understanding the meaning of any seemingly strange results in the data.

Quality assessment often reveals unexpected patterns and challenges prior assumptions. As we have already seen in numerous examples, producing simple descriptive statistics (distributions, frequencies, totals etc.) enables us to gain a more detailed understanding of the metadata, raising questions which can be put to the data provider.

Such interactions are invaluable: when working with administrative data, we have few real opportunities to act retroactively on the data. Phase 1, and particularly the quality assessment step, is **the** moment when we can exchange directly with administrative agencies. This is also an opportunity to strike up a constructive dialogue, in the long-term interests of administrative processes and the data available to statisticians.

- ***A ‘macro’ feedback loop***

In our first chapter (section 1.c in particular) we demonstrated the importance of verification in the survey process, looking closely at the concept of *data editing*. The situation we described was an example of a feedback loop, combined with automated preliminary checks capable of detecting *anomalies*. However, this process led to modifications to the microdata, at the level of individual statistical units: thanks to the work of the person conducting the survey, conversing directly with the respondent (an immediate, or ‘live’, feedback loop), or else thanks to the subsequent intervention of a survey clerk, contacting respondents with follow-up questions (*manual editing*).

An example of an anomaly which can be resolved using a feedback loop: for a given company, the value of the variable Primary activity, 47.21A, is outside the acceptable range (in this case the NAF, the French classification of economic activities). We can contact the company to ask why this might be, and we may find out that it is simply a data entry error. The feedback loop was thus triggered by a piece of microdata, and the result will be a modification to that microdata.

When working with administrative data, the quality assessment phase will also allow us to detect surprising or abnormal situations, and to subject them to critical analysis, but both the trigger and the outcome of the feedback loop are to be found at the macro level.¹⁰⁷

For example, consider this anomaly: 5% of businesses listed in a data file have 47.21A declared as their primary activity. We contact the data provider to ask why. Perhaps the administrative agency has voluntarily introduced a new activity code, for their own purposes? There is no question of asking the individual units (companies, in this case), and the result of the feedback may be something along the lines of a change to the accompanying documentation.

Macro-anomalies come in many guises (cf. 4.d): frequent domain anomalies, undercoverage, implausible value distributions, other coverage problems (volume and or types of objects), units outside the expected scope, time frames not respected and so on.

¹⁰⁷ Made possible by adopting the preferred terminology of the administrative agency/data provider.

Feedback processes, including discussions with administrative data providers, therefore seek explanations for these anomalies. They can be divided into three categories:

1. Misunderstandings on the part of the statistical office,
2. Errors or shortcomings in the metadata (and documentation),
3. Errors or shortcomings in the data shared.

The first of these issues can be solved by means of explanation, and it is down to the statisticians to improve their own documentation. Returning to an earlier example: the anomaly “81st floor” does not stem from a problem with the shared data: they simply lacked the important information that the code 81 is used to represent a basement floor, or level -1.

In the second scenario, we may identify a value which we would not expect to see in a particular domain (for example an activity code), or discover that the scope is a little broader than we had been led to expect. The feedback loop will thus yield new versions of the supporting documentation (sometimes just a minor update), with no need for further data transmission. This is not far removed from the first scenario.

This leaves us with the third scenario: the documentation is not the issue, so there must be a problem with the data as they have been delivered. The fact that the scope is incomplete may indicate that delivery is not in fact complete (some counties may be missing, for example). The presence of duplicate entries may be indicative of inconsistencies in the information system, requiring in-depth work on the part of the administrative body. The same applies in cases involving abnormal value distributions across a domain. This may require additional data deliveries, “superseding and replacing” previous deliveries.

It is worth noting at this juncture that it is not always possible to make all of the necessary modifications within the desired time frame.¹⁰⁸

Any modifications made will not apply to individual data units; they will come at a more systemic level, with the indirect effect of improving the administrative data, which will be very useful if this is a recurring exercise. Boydens (1999) thus proposed a positive approach to using the anomalies found in Belgian social security data, subsequently expanded to other contexts (Boydens 2018). Instead of “feedback loop”, Boydens prefers the term “back tracking”, inspired by Redman’s (1997) “data tracking”, focusing on the efforts made by administrative agencies to continuously improve the quality of their data, and thus differing slightly from our own analytical framework, which is focused on statistical uses of public data.

To wrap up this section on feedback, let us not forget that it may be a recurring exercise: data quality assessment is a progressive, iterative operation which spans the whole process of exploring and acquiring administrative data, and beyond.

4.f. Repeat and multiple uses

- *Evaluating the quality of sources intended for repeat use*

While quality assessment is indispensable when first acquiring a new data source, it is every bit as important when accessing data on a repeated basis. Once the exploratory phase is completed, statisticians are familiar with the source and ready to begin using it. Proper documentation of their first experiments with the data is essential, allowing for the accumulation and transmission of knowledge regarding the source.

¹⁰⁸ In contexts involving repeated data acquisition, detecting anomalies may lead to changes to subsequent deliveries, thus necessitating corrective statistical measures for the existing data.

When working on an annual basis, for example, the data acquisition process should be simpler in the second year: the task of becoming familiar with the administrative context has already been completed, and the preparation and data-receipt and quality-assessment procedures should be tried-and-tested, with sufficient documents and tools available. Of course, the data extraction and transmission processes will still need to be checked, along with the usual compliance and plausibility checks and the familiar feedback loops. Nonetheless, the initial investment should still be paying off.

However, new questions are sure to emerge regarding potential developments: has the applicable legislation changed? Have any of the classification systems or concepts been modified? Is there a new software system in play, or new formats?

Answering these questions requires some detailed investigation. This begins with **monitoring legislative developments**, always an essential parameter. It also demands regular exchanges with the data provider, and special attention to the metadata provided with sources, but also – and this is a point which is frequently overlooked – a particular focus on any changes in the distributions of important variables. Comparing multiple time-specific iterations of the same dataset can allow us to introduce temporal consistency checks into the quality assessment process, making the process more richly detailed... and more complex.

The need to satisfy quality criteria may evolve as statistical use of the source becomes more mature: when acquiring data for the first time, evaluation is required to determine the quality of the source and its capacity to satisfy statistical needs in the short term; however, this evaluation may also identify changes required to ensure that these needs are satisfied in the longer term. Quality is not invariable over time, and we must be able to spot the changes: the coverage ratio of a source may improve, the time it takes to obtain a source may be reduced, or, at the other end of the scale, a source which is scheduled to disappear may see a drop-off in quality.

Example: the abolition of the council tax in France led to a deterioration in the coverage of the connection between fiscal households and their homes. We had details for 37 million households and their home addresses in 2022, 27 million in 2023 and fewer than 7 million in 2024. Remedial measures were thus put in place to adapt the information system. This deterioration in the quality of the data allowing us to locate individuals was expected, and preparations were made to ensure that the data for these years remained usable, with the help of the GMBI source.

The fact that quality assessment becomes a repeated exercise has another noteworthy consequence: we must ensure that we are able to trace and reproduce any data processing actions (Pérignon et al 2019). This is a matter of quality development (legibility, version management), which also highlights the growing importance of proper documentation.

Finally, provided that certain changes are unavoidable (changes to the legislation, for example), we must take care not to end up with an automated process (of the pipeline variety) which is too rigid, too difficult to update. The process must remain flexible enough that we can modify its component elements, adding new variables, domains, quality indicators etc. What we have here is essentially another form of feedback loop, this time an intertemporal loop, concerned not with the process of delivery, but instead with the acquisition process itself.

- *Is it possible to evaluate the quality of a source for multiple uses?*

If quality depends on intended use, then assessing the quality of a source for multiple uses may seem paradoxical: different uses may have diametrically opposed objectives (the classic trade-off between data timeliness and comprehensiveness), or the need to consider different uses may result

in overly complex analyses. Nevertheless, when a source is frequently used in different production processes, it may be advantageous to pool some of the quality assessment work, and particularly to share the results of that work (as witnessed by the collaboration between DREES and INSEE regarding the PASRAU source): sharing new knowledge, understanding the source, even automating the production of quantitative indicators. This is particularly relevant in contexts where data reception is centralised for multiple users.

Pooling can take multiple forms, from exchanges within a working group to sharing the results of quality assessment analyses focusing on different domains. When a new potential use is identified, it may be necessary to perform additional analyses on a source which has already undergone quality assessment in the past.

Example: data from the land register previously used to locate residential properties underwent further analyses when they were used by the Fidelimmo¹⁰⁹ project (André & Meslin, 2022). This project was interested in the variable ‘floorspace’, which was irrelevant to previous users, and had thus not been subjected to quality assessment .

4.g. Outcomes of the acquisition phase, and moving to the next phase

The acquisition phase is a particularly uncomfortable time for statisticians, since it has little to do with their usual skills and academic backgrounds. It requires them to become immersed in an unfamiliar administrative domain, with its own terminology, semantics, operational processes and objectives.

It also involves a significant amount of logistical planning, organizing delivery mechanisms, checking and monitoring those mechanisms, and drawing up contracts to ensure that everything runs smoothly. There is also a certain amount of negotiation involved, as updates and iterations take shape with the help of feedback loops.

The outcomes of this phase are as follows:

- Data files which, by definition, are frozen;
- Metadata, accompanying documentation enabling us to understand the meaning of the data and use them more easily: definitions of the objects, domains, scope, variables and time frames;
- A data sharing agreement;¹¹⁰
- An analysis of the quality of the data files received: we may not necessarily have files of outstanding quality, but rather files “whose weaknesses are known to us” (incomplete and non-compliant data etc.).

All of the work done during this phase is conducted **in the preferred language of the administrative entity**, using their objects, variables etc., and the statistical team may in **no way modify the data**.

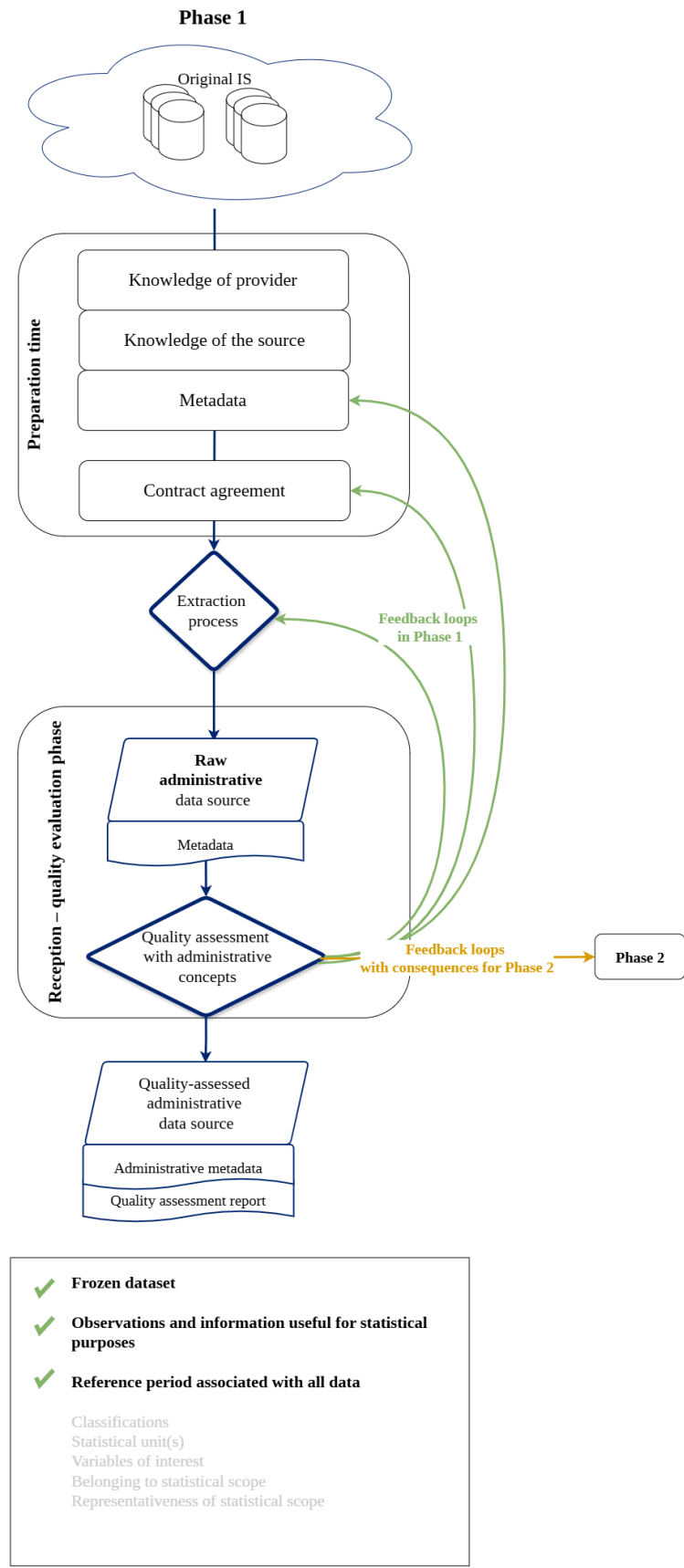
With regard to the feedback loops established with administrative partners, there is one more thing that we should note: in practice, there may well turn out to be no feedback. The administrative entity may not send any more files, documents or explanations, perhaps because they simply do not have the time. Hence the vital importance of quality assessment: this process enables us to identify the weak points of the source, weak points which will need to be taken into account, or even

¹⁰⁹ Project that aims to build a new statistical database, using administrative data sources and systematically combining information on households and on the real estate assets they own

¹¹⁰ This is an *output* when acquiring the data for the first time. In subsequent acquisition periods, this agreement will still be in place but may require amendments.

“remedied”, in the subsequent phases. But to do this, we must first pass over into an environment better suited to statistical work: this is the focus of Phase 2, the transformation phase.

Figure 4 - Phase 1 - Acquisition of administrative data



5. The transformation phase: making the transition from the administrative realm to the statistical realm

The very fact that the data produced by Phase 1 belong to the administrative realm constitutes a limitation. We cannot use these data in this form: they must be **transformed** into a language and framework better suited to statisticians. To borrow Courmont's (2021) terminology, we must thus "detach" them from their native realm and "re-attach" them to the statistical realm. This amounts to a change of register, a projection of the administrative realm into our chosen statistical dimensions.

The objective of **Phase 2, known as the transformation phase**, is to obtain a dataset suitable for statistical use, provided that the data correspond to the units and variables identified when we determined our statistical requirements.

This transformation requires us to perform a series of operations on the files, operations which are often planned out well in advance.¹¹¹ The data obtained at the conclusion of this phase will thus be similar in "status" to survey data, which is essential for the subsequent phase.

In this chapter, we begin by describing the different operations which make up the transformation process. We then consider the task of replicating this chain of operations, and find that quality assessment (integrating a feedback loop) is still necessary, albeit of a different nature.

5.a. Principles

To move from the administrative to the statistical, we need to make the transition from one analytical framework to another. We want the results of this transformation to:

- Be based on a clear understanding of the objects, scope, variables, domains and time frames of the data files, using the semantic codes of the administrative agency;
- Allow us to better comprehend the statistical units, scope, variables, categories and time frames chosen for statistical purposes. We thus systematically refer to these different dimensions, while bearing in mind that the question of time frames was largely settled during Phase 1.

The transformation process must be designed with these objectives in mind, with several points which demand particular vigilance:

1. The resulting data could be of use to other statistical offices, for other purposes. If so, the demands of data sharing should be borne in mind from the outset (this is an important issue when working with the DSN, for example), taking different potential uses into consideration;
2. Depending on the choices we make, the transformation process could render the data unsuitable for other uses. The fact that a process might lead us to delete some information (excluding certain records) and aggregate other entries (combining multiple records to form new statistical units), means that some information is lost in the post-transformation data, which may preclude other uses.
3. Transformation is not merely a program: it is an object of study in its own right and must be understandable, adaptable and properly maintained

¹¹¹ Which may have implications for the design of Phase 1.

5.b. Principal operations in the transformation phase

The process can be broken down into steps, transforming the different characteristics of the data: object, scope, attribute (or concept), domain and even value. More often than not these steps are deterministic, and not always reversible. Cotton and Haag (2023) propose breaking the process down into six steps.¹¹²

- *From administrative object to statistical unit*

Restructuring the data into statistical units

Administrative data are concerned with objects of interest to the administrative body, connected to a process and its objectives. These objects may be complex, and different objects are often interconnected (individuals, establishments, events etc.).

Official statistics often provide data on **individuals, households, businesses, legal entities or establishments**. They also make use of event-specific data, usually very closely connected with these “familiar” units (births, deaths, business creations, transfers, restructuring operations etc.).

The transition from administrative objects to statistical units is often a complex one, and requires a sound knowledge of the data. A single data source can be used to construct multiple datasets, based on different statistical units.

For example, DSN data, which can be viewed as *payroll record line items*, enable us to construct data for individual *jobs*, but also for individual *employees*.

We sometimes need to use multiple objects, and the connections between them, to arrive at a statistical unit.

In the tax data, the connections between residential addresses and their occupants are used to define households. Another example: in business statistics, shared ownership connections between legal entities allow us to identify groups of companies.

The transition from objects to units is not necessarily bijective: objects may be split or merged, or we may choose to focus exclusively on certain connections. In cases where statistical units are made up of multiple objects, we end up with a form of duplicate removal which is not “technical”, but informed by concepts.

Example: DREES uses the DRM¹¹³ data to track the monthly income of certain individuals. The source contains lines corresponding to income and benefits *transferred from a paying party to a beneficiary*. The statisticians at DREES are interested in knowing whether these payer-beneficiary links are new or longstanding, allowing them to track income and benefits over time. But, in practice, some administrative agencies create a new relationship for each new payment, i.e. generating a new object for the “service” instead of modifying an existing object. This means that work is required to identify entries pertaining to the same statistical unit, i.e. the same pairing of payer and beneficiary, and involving payments.

¹¹² Namely: filtering useful entries, renaming, restructuring data into statistical units, recoding, calculating derived variables, pseudonymisation.

¹¹³ DRM = Dispositif de Revenus Mensuels / Monthly Income System.

See https://net-entreprises.custhelp.com/app/answers/detail/a_id/2362/~/donn%C3%A9es-dsn-et-pasrau-par-le-dispositif-de-ressources-mensuelles-%28drm%29.

For a discussion of its use in the social sphere, see(Puche 2022).

The fact that the statistical unit and the administrative object are very different things may have consequences when it comes to identifying units.

The land register, for example, contains data pertaining to property transactions. In some cases, the administrative object is the property changing hands during the transaction, listed alongside the names of the individuals involved in the transaction; however, the details provided regarding these individuals may not always be precise enough to allow for their identification. This makes it difficult to establish statistical units corresponding to “individuals” (André, Meslin 2022).

Pseudonymising statistical units

Making the switch from objects to statistical units is also a matter of ensuring that the latter have the required properties for statistical use: for data pertaining to individuals, one key property is that personal details should not be accessible; i.e. it must not be possible to connect individual persons to the data held about them. There are numerous methods available for protecting data confidentiality (Matthews, Harel 2011). They are known collectively as Privacy-Enhancing Technologies (PET), adopting a comprehensive approach to this task and incorporating recent developments in the field of official statistics (Ricciato 2024).

Pseudonymisation is one possible solution. It consists of removing, for each individual contained in our files, the direct identifying information along with any identifying features. It is important to be clear about the distinction between pseudonymisation and anonymisation (Esayas 2015). In the words of France’s National Commission for Information Technologies and Civil Liberties (CNIL), “Anonymisation is a process which uses multiple techniques to make it practically impossible to identify a person, by any means whatsoever, and irreversibly.”

In other words, anonymisation is much more demanding, and may require us to deliberately discard a large quantity of information.

Pseudonymisation, however, is a “relatively” simple operation (at least conceptually), and can be done automatically: it can easily be slotted into the series of operations which comprise Phase 2 of the transformation process.¹¹⁴ In France’s official statistics ecosystem, the CSNS¹¹⁵ (for *Code statistique non signifiant*, or non-significant statistical code; see Benichou, Espinasse, Gilles 2023), offers a way of implementing this operation.

¹¹⁴ See also (Girard-Chanudet 2023), who offers an entirely original insight on the automation of pseudonymisation, in a context far removed from the realm of official statistics.

¹¹⁵ The Non-Significant Statistical Code (CSNS) was defined by the Law for a Digital Republic of 2016 in order to allow the implementation of file matching for statistical purposes without using the NIR (or national insurance number), thus ensuring a high level of personal data protection.

- *From administrative scope to statistical scope*

When discussing scope, we must begin by clearly defining what scope we mean: it may be the scope of the administrative data themselves, or the scope of our statistical interest. During the transformation phase, we move from one world to the other, bearing in mind that the administrative scope is imposed by the source. As such, there is no guarantee that it will be sufficient to cover our desired statistical scope. During the transformation phase, we focus primarily on eliminating entries which lie beyond the scope of our interest. In other words, this phase is more concerned with tackling *overcoverage* rather than *undercoverage*: there is no way of filling in any gaps in the scope (that is a matter for Phase 3).¹¹⁶

Removing duplicates

Resolving issues of overcoverage is primarily a matter of deleting **duplicates**: ideally, any duplicate entries stemming from the administrative context will have been detected in Phase 1. Nonetheless, it is possible that they were not dealt with by the administrative entity during the feedback loop phase, or indeed there may not have been any feedback loop at all. A round of “technical” duplicate removal is required, detecting duplicates of administrative objects on the basis of their identifying features, using the available tools and methods: Malherbe (2023) offers a comprehensive overview.

This is very different from the “conceptual” duplicate removal which may be required further down the line, when making the transition from objects to units (for example, a single business may submit multiple tax declarations).

Filtering for statistical scope

The goal here is to use the available categories to eliminate anything which is not relevant to our stated statistical scope.

For example, we may decide to exclude units from certain sectors of the economy (the non-market sectors, for example), or individuals present in the administrative data who do not belong to the resident population, or properties not used for residential purposes... it all depends on our intended use for the data.

Another example: every month, the CNAF sends INSEE a file containing details of individuals who have received family benefits in recent years. But if we want to focus solely on those receiving benefits in the past month, we must filter the file with a specific setting for the variable allowing us to identify this “core” group.

Filters are defined based on the variables available in the source. They are calculated using the entries recorded in the data, on the basis of existing information, and may lead us to select objects, units, or only certain components of objects or units.

For example, the tax data contain information on residential properties. This concept is not strictly synonymous with the notion of a “home”. Certain choices are therefore required in order to switch from one concept to the other. In this particular case, we must discount car parks, outbuildings etc., which are classed as residential properties but are not “homes”.

These filters involve varying degrees of complexity. It is essential to explicitly define the filters and their relative positioning within the process: the order in which processing operations are performed

¹¹⁶ On this subject, see Wallgren & Wallgren (2016).

may affect the results. It is also worth noting that by precisely defining a filter we presume that the filter variable involved is accurate.

The quality of filtering variables needs to be meticulously evaluated: errors of scope may have an outsized impact on the quality of the resulting statistics, and are frequently to blame for the problems encountered by statisticians. The variation of these variables over time must also be closely monitored, with a clear understanding of the administrative process and applicable legislation, and any changes made in the implementation of the process.

A recent administrative change provides an instructive example: the new obligation for “anybody applying for the RSA¹¹⁷ benefit, young people supported by local agencies and people with disabilities seeking specialist support” to sign up with France Travail¹¹⁸. This constitutes a change to the definition of “jobseekers”, with major implications for the corresponding statistics, leading inexorably to a significant increase which is hard to interpret. The situation is so problematic that a working group (led by the CNIS) has been formed to address the subject. The group has proposed a series of categories which “mirror” the new administrative categories: persons registered “on social programmes” (Cat. F) and RSA applicants receiving guidance (Cat. G). With the help of these categories it becomes possible to reconstruct the scope appropriately, but close monitoring is required.¹¹⁹

Changes of scope resulting from administrative decisions may be identified in advance, as in the preceding example, allowing us to anticipate the consequences. The opposite may also be true, and it may be substantial variations in observed volumes that alert us to changes in the scope of the source, and thus the underlying administrative measurement.

Example: the introduction of pay-as-you-earn tax deductions in France changed the population found in the fiscal data sources. This change was detected when we observed the consequences in terms of the number of individuals contained in the sampling frames constructed by INSEE. The complexity of the processing operations undergone by fiscal data makes it difficult to analyse volume variations in these data.

Another example: in recent years there have been some spectacular increases in the membership numbers of French sporting federations, which are in fact due to changes in the way these federations operate. Membership numbers now include those who practice sports on a non-competitive basis. For example, the number of membership cards issued by the French canoe-kayak federation virtually doubled between 2016 and 2019, as a result of such changes (Vicard 2023).

- ***From administrative variables to statistical variables***

We often find that the variables contained in administrative sources are not exactly as we would like them to be, but that they nonetheless suffice to construct approximations of the variables of interest to us. A lot of initial work is thus required to understand the meaning of the data (administrative and

117 The Active Solidarity Income (RSA for Revenu de Solidarité Active) guarantees a minimum level of income for people without resources.

118 France Travail (English: France Employment) is a French governmental agency which registers unemployed people, helps them find jobs and provides them with financial aid.

119 Cf. the summary of the CNIS working group’s report “Consequences of the implementation of the ‘Full Employment’ act for unemployment statistics”: <https://www.cnis.fr/wp-content/uploads/2024/09/rapport-cn166.pdf>

statistical), before we arrive at an operational calculation formula and start attaching names to things.

Calculating the derived variables

This is the term used to describe variables (in this case, statistical variables) calculated using other variables (here, administrative variables) by means of a calculation formula.¹²⁰

Examples: variables used to reconstruct wages based on the DSN. We may introduce rules along the lines of: use uncapped gross earnings, if that is not available then use the uncapped gross tax base, if that is not available then use the CSG base¹²¹.

Another example: Van Delden et al. (2016) reconstructed turnover figures using several variables contained in Dutch VAT declarations, and had their calculations checked by experts in the field.

As illustrated by the example above, the calculation formula does not always take the form $z = x + y$. It may also have to contend with the presence or absence of certain information, constructing a variable which is not troubled by such lacunae. The aim is to produce the best possible proxy for a statistical variable using the administrative variables at our disposal. In such cases, it is vitally important to keep a record of the exact calculation formula used (which is, in fact, a form of metadata).

N.B. derived variables may still differ from the statistical variables we are aiming for.

In the Dutch example cited above: calculated for sectors of activity constructed at a fairly fine level of detail (324 group), total turnover derived from Dutch VAT declarations does in fact differ from the turnover data derived from survey responses. These conceptual differences can be very subtle.

The composition of statistical units also has ramifications for the calculation of variables linked to objects. When working with administrative declarations, we may choose to take the declaring parties as our statistical units, aggregating multiple initial entries (the declaring party can send several declarations) into a single unit with the help of aggregation functions (depending on the variables, they might be sum totals, maximum values or most recent values).

An example relating to structural business statistics: moving from the object “tax declaration” to the statistical unit “legal entity” requires us to aggregate the declarations. The calculation methods used will differ depending on the nature of the variables: we may simply add up flow variables (turnover), but for closing inventory the most recent value is the one we want. We must also take into consideration the fact that fiscal data are relevant to accounting periods, which may not correspond to the calendar year: the calculation formulae may turn out to be complex.

¹²⁰ See for example the index of variables used in the 2021 census in Northern Ireland, where this distinction is clearly explained: <https://niopa.qub.ac.uk/bitstream/NIOPA/15777/1/census-2021-variables-index.pdf>

¹²¹ The contribution sociale généralisée (CSG) and the contribution au remboursement de la dette sociale (CRDS) are taxes designed to finance social protection. The remuneration base subject to tax may differ from the employee's actual remuneration. DSN contains different remuneration bases.

Renaming

Changing the name of a variable is a reversible operation which has no consequences for the data themselves, but does affect the metadata associated with the data. As renaming is a simple, inexpensive operation, there is a certain tendency to underestimate its importance, or to overlook it entirely. But it is important in order to ensure that the concepts employed are well understood, durable and widely shared.

Records must be kept of all renaming operations, documenting the data for end users. Renaming has an impact on an important dimension of quality: legibility.

For example, fiscal data use names which are entirely incomprehensible to non-specialists, as witnessed by the examples cited by Cotton & Haag (2023). Name, first name, gender and date of birth are written, respectively, as LNCON, LNCOPF, LCCOT and DNCO.

Rigorous and clear renaming allows us to save time on subsequent operations, since we now know what we are talking about. Renaming may concern either the variables or the values of a domain.

- ***From administrative domains to statistical categories (or domains)***

In order to illustrate the process of domain transformation, consider the example of an initial variable for which the initial domain is {a, b, c, d, e, f}, which we want to simplify: for example, by using the new values 1, 2, 3, where 1 corresponds to {a, b, f}, 2 corresponds to {c}, and 3 corresponds to {d, e}.

In some respects, the situation here is similar to that we encounter with derived variables: behind the domain change lies the calculation of a derived variable, whose values are taken from a list. We call this coding, or recoding if the output is a list of codes. The value domain of the input may be a list, as in our initial example, but it may also be quantitative data (e.g. moving from specific ages to age groups). It is also important to note that value calculations in the target domain may involve multiple inputs, with complex calculation rules.

These rules must also take account of missing values and their denomination, both upstream and downstream. Upstream, thanks to Phase 1 we know how “missing” values are recorded in the administrative data (for example, 999, 9999, -1 etc.). Downstream, on the statistical side of things, we use an agreed form of notation (for example ‘NA’), accompanied by decision-making rules. Nonetheless, it is important to note that during the transformation phase we do not process missing data and we do not engage in imputation (that is a task for Phase 3): we simply establish and apply notational conventions and attribution rules based on the administrative input.

Domain transformations must also take account of changes in management practices, as illustrated above by the new rules for registering with France Travail.

5.c. A technical process, which we must be capable of replicating

Transformation also involves some technical challenges, which may stem from the volume of data involved, the frequency of deliveries, the format etc. The complexity of the originating information system, particularly the connections between objects, is reflected in the statistical data available to statisticians.

When making technical implementation decisions, we must take the intended uses of the data into consideration:

- We must strike a balance between the frequency of data uploads and the stability of the database,

Example: the daily deliveries of PASRAU data cannot be channelled directly into the data available to statisticians. In order to guarantee a certain degree of stability, the databases available to statisticians are updated on a monthly basis;

- the structure of the data and the architecture of the databases must be compatible with these uses, making it easy to filter information and establish internal connections.

For regular data sources, we need to put in place an automated production process, a series of transformations often known as a *pipeline*.¹²² As we saw in Phase 1, this pipeline must be sufficiently flexible to allow for changes to be introduced without too much difficulty. It must also guarantee, up to a point, replicability (National Academies 2019),¹²³ something which has consequences for data storage before transformation, the traceability of data throughout their life cycle (also known as *data lineage*), and, often, the capacity to handle large volumes. Ideally, the degree of replicability should be high for those production processes which have serious consequences for the information system.

But we must above all bear in mind that transformation, as an automated series of operations, leads to the production of a new source, a source which exists in the statistical realm: it will contain individual data pertaining to the statistical units of interest to statisticians, with their chosen scope, variables and categories. The problem is that, just as at the end of Phase 1, this new source will contain certain imperfections. Another round of quality assessment is required.

5.d. Evaluating the quality of transformed data

As with the quality assessment conducted at the end of Phase 1, we must now assess the quality of the data we receive (produced by the transformation process, this time) in order to detect any anomalies, then put a feedback loop in place so that these anomalies are duly accounted for. Although the principle remains the same, these quality assessment and feedback processes are nonetheless very different from those used in Phase 1.

- **Principles**

Just as at the end of Phase 1, the aim is to detect any problems and identify any suspicious data. This time, however, we are operating in the statistical realm, a world which is familiar to us, about which we have ample information.

The quality assessment conducted at the end of Phase 2 is an intermediate, hybrid process because it evaluates the transition from one realm to another; although we have passed over into the statistical realm, we have not yet completely left the administrative realm behind. This evaluation is hybrid to the extent that it focuses on both individual data and aggregates. It thus operates in a particular manner, leading to consequences of two distinct types: it triggers a feedback loop when statistical anomalies are detected, and facilitates Phase 3 via the systematic spotting of individual anomalies.

¹²² On this subject, see (Cotton, Haag 2023), already cited above.

¹²³ It is interesting to note that, among the committees formed for the purpose of drafting this book, there is indeed a 'Committee on National Statistics'.

Detecting statistical anomalies

Genuine statistical quality assessment is a matter for Phase 3, when we have “all” of the data at our disposal, imputed or not. For now, in Phase 2, we will limit ourselves to a few **broad-brush statistical evaluations**. The missing data and units have not yet been processed, so we cannot construct valid statistics at this juncture. We must therefore simply check that the data are in the correct order of magnitude, verifying:

- Consistency with the totals we already know, for example the number of households in Fideli (Lamarche & Lollivier 2021);
- Changes in the totals, allowing us to spot changes in scope (cf. the France Travail example);
- The extent to which the scope is covered, particularly when the administrative and statistical scopes are comparable: this gives us an estimate of coverage problems;
- We might also potentially check the plausibility of a few distributions, using other, known distributions (from census or survey data).

In conducting such comparisons, we are now operating entirely outside the administrative realm (we can use any other source which we consider to be reliable and comparable, at least for certain variables). *We can now compare variables and statistical units.*

This means that we are now able to count the number of households in Fidéli (the statistical product), rather than the number of fiscal households present in the FIP¹²⁴ source (original administrative data). By the same token, we can now compare net earnings at full-time equivalent, rather than the DSN declaration lines.

Such exercises are conducted systematically during the quality assessment process, but may also be used on a more occasional basis in relation to new uses of existing administrative sources.

An example of this comes from the DSN data, which have already been used for several years to produce annual statistics on salaries, not least the ‘all-employee database’. The DSN data have since undergone quality assessment for a new use, and in 2024 this made it possible to reduce the number of questionnaires used for the Labour Cost Survey by a third. This usage-specific evaluation was conducted independently of the process which produces the all-employee database.

Identifying individual anomalies

The aim here is to “scan” the file in the hope of automatically detecting missing values (partial nonresponses) and aberrant values, using the *data profiling* techniques detailed above (Olson 2003). Missing values can be spotted simply enough because we know the conventional notation used to denote “missing”. Spotting abnormal data is a more complex affair, and requires us to define some rules in advance: for example, if a company’s turnover / headcount is above a certain threshold, for a given sector of activity, we can regard it as being abnormal.

This step leaves us with a complete data file including both data (microdata) + micro-metadata identifying all missing or abnormal values.

124 Fichier d’imposition des personnes (Tax File for Individuals)

- ***The quality assessment conducted in Phase 2 is no substitute for the assessment required in Phase 1***

At this point, it is worth noting that certain problems in the source cannot be detected in Phase 1 (they can only be detected by means of statistical comparisons) whereas others could have been, but might have escaped the notice of statisticians due to shortcomings in the process of quality assessment in Phase 1, or simply due to a lack of time: in many statistical processes, data reception and integration take place within a very tight period.

When acquiring an administrative source for the first time, the data are still largely unfamiliar, and the transformation process still needs to be pieced together. In such circumstances, we naturally conduct quality assessment at the end of Phase 1, assessing the quality of the administrative data with reference to their own context, concepts, objects etc.

However, when working with administrative sources on a regular basis, we may be tempted to think that the quality assessment at the end of Phase 2 will suffice, and thus dispense with the evaluation in Phase 1. This is all the more tempting when the process of data integration is bulky, elaborate, old etc.

Although it may seem rational enough, slimming down quality assessment in this manner can be problematic on several fronts:

- It makes it harder to identify the origin of problems: does a given problem stem from the data, from the transmission process or from the transformation process? At the very least, if there is no dedicated quality assessment at the end of Phase 1, it is important to keep track of data volumes (delivered, integrated, transformed) or to have some general descriptive statistics regarding the data pre-transformation;
- It makes exchanges with the administrative data provider more complex and costly;
- It may “delay” the detection of changes in the administrative data, or of new quality issues, especially if quality assessment is conducted later on in the process, sometimes even after the statistical processing operations which make up Phase 3. Such delays can be crippling when producing economic outlook statistics, where publication deadlines are very tight.

5.e. A feedback loop operating on two levels

When quality assessment detects problems in the data, there are only two possible culprits: the source, derived from Phase 1, or the transformation process. The two are intertwined and difficult to separate. If there really is a problem with the source, we should ideally be able to spot it at the end of the acquisition phase, putting in place a feedback loop with the administrative entity at this juncture, ahead of transformation. But this is no easy task.¹²⁵

- ***The preferred feedback loop: focused on the transformation itself***

We thus begin by devoting a feedback loop to the transformation process itself (with the implicit assumption that the source is of good quality): if anomalies are present, they must be the result of bugs in the transformation operation, and we must find those bugs. Transformation must thus be replicable: if it includes manual steps, or undocumented automatic operations, then it becomes very difficult to make changes.

¹²⁵ It requires us to keep a record of the data as they were at the end of Phase 1, which can take up a lot of space. We also need to repeat the quality assessment from the end of Phase 1 ... and succeed in spotting the problems.

Anomalies can come from a number of sources:

- Programming errors;
- Errors in the interpretation of the administrative data (definition of variable or value in a given domain);
- Incorrect formulae used for calculations: the formulae used to calculate filters or derived variables must be very accurate, and cover all eventualities;
- subtleties in the formation of statistical units;
- and above all: changes to administrative management tools, legislation or administrative processes, leading to changes in the concepts, domains etc.

Example: as part of the process employed to assess the quality of the DRM data, DREES has created a sub-sample which allows for the most comprehensive possible diagnosis of data quality problems. The goal is to establish criteria for detecting such problems, and the corrections required, while also specifying the criteria for forming statistical units. This sample is also used to verify processing operations. Once the data transformation process has been defined in writing, it can be applied to this test sample, allowing the statisticians to verify that the results match their expectations.

Errors are often induced by small changes to administrative procedures, sometimes so small that they are hard to see. This is one of the major drawbacks of working with administrative data for statistical purposes, highlighting once again the need for a thorough understanding of the administrative realm whence such data originate.

All things considered, putting together a data pipeline is a very demanding operation, a feat of precision mechanics with no room for even the slightest error, requiring advanced skills in data engineering (Schweinfest, Jansen 2021).

- ***Feedback loop to Phase 1, a more delicate task***

Although detecting problems in the source is not an ideal scenario in Phase 2, it can happen for various reasons, not least because comparisons with other statistical sources allow us to obtain a more accurate estimation of undercoverage, along with a clearer picture of the values when it is possible to compare individual data.

Discussing the matter directly with the data provider is always useful, but we now encounter an additional layer of complexity: since the data have been transformed, problems will most likely be identified and expressed in statistical language, dealing in concepts which are different from those found in the administrative source. *This is a major source of misunderstanding between the administrative and statistical worlds.* Considerable effort is thus required to reformulate any anomalies detected using vocabulary and concepts which are comprehensible to the administrative entity which provided the data.

Once again, caution is required because various changes made to the administrative source may render certain parts of the process obsolete.

Example: the consequences of the transition to a pay-as-you-earn system for income tax collection in France. Henceforth, the entity responsible for providing salary information is no longer the individual taxpayer, but their employer. This change to the administrative process has led to changes in the coverage ratio of the data source, with improved coverage since the data now include income which individuals may have failed to declare in the past. In the first year after this change, the statistics produced by the Fideli process contained a

million new individuals (!). The transition from “taxpayers” to “persons residing in France” had not yet been fully integrated in the statistical process.

Diagnosing the reason for this huge gap was not easy. From a statistical perspective, the change to the administrative process served to correct an existing undercoverage problem (some income was previously not being declared by individuals residing in France), but it also exacerbated the problem of overcoverage (for example the earnings of individuals not resident in France, such as seasonal workers).

5.f. Outcomes of the transformation phase

Less awkward than Phase 1, Phase 2 of the transformation process is more within the wheelhouse of official statisticians, not least with regard to the IT requirements. The aim of this process is to define a series of operations capable of translating data into a familiar framework.

The desired outcomes of this phase are as follows:

- “raw statistical” data,¹²⁶
- accompanying metadata,
- the transformation code,
- the results of quality assessment.

This may all seem a bit strict, and a few comments on each point are perhaps in order.

Why keep the “raw statistical” **data**? Because if we encounter a problem in the subsequent statistical processing operations, it will still be possible to modify the operations and repeat them using the raw statistical data. Keeping these data is therefore essential in order to ensure the replicability of processing operations. Moreover, unlike the files transmitted by data providers, raw statistical data can be archived by the statistical office: they thus have the same status as raw survey data.

The **metadata** (definition, domain characterisation etc.) are gradually accumulated as the transformation process takes shape. We need simply take care to save them properly, in order to retain control over the results, but also for potential future uses or data sharing agreements.

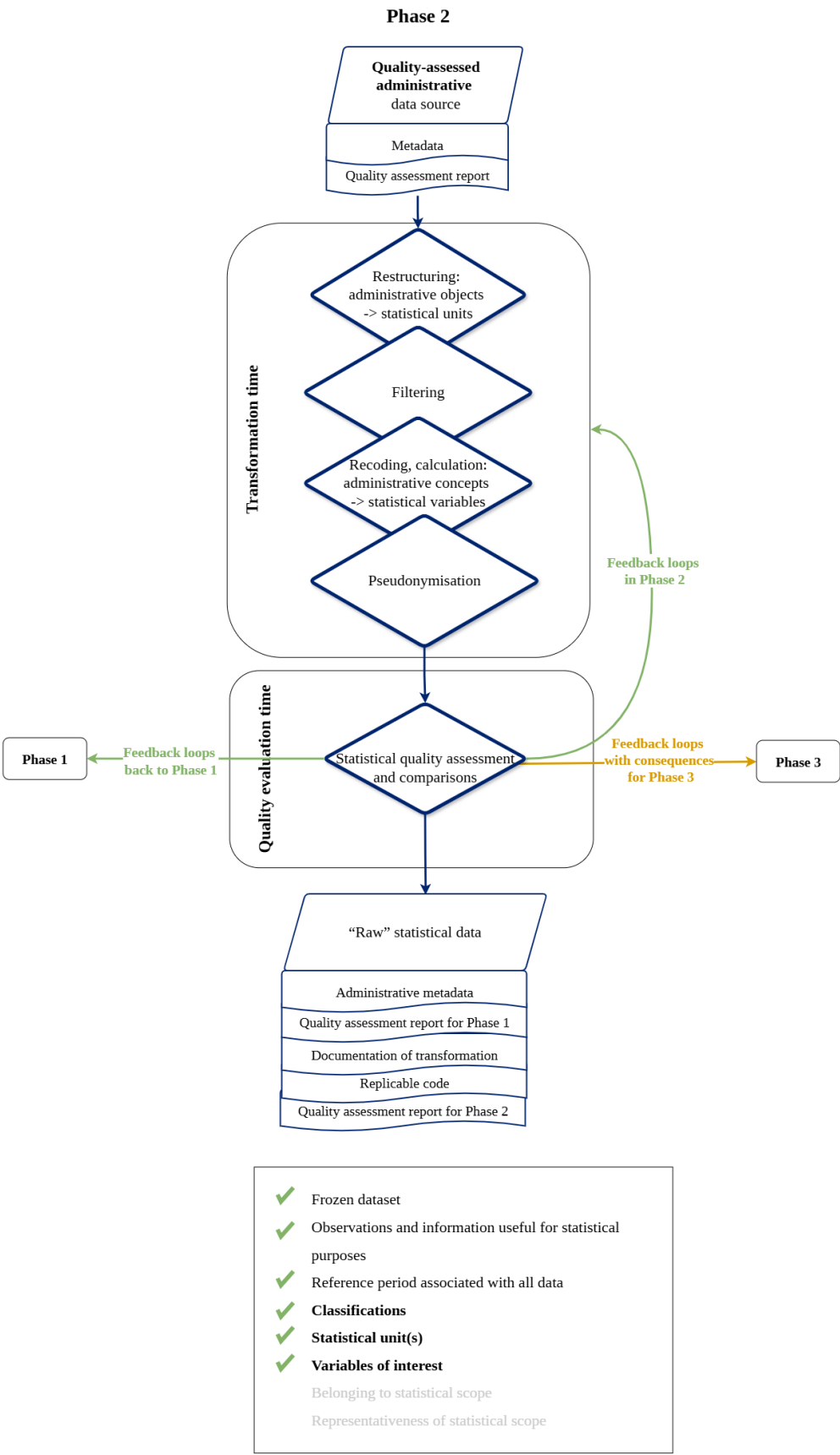
As far as possible, **transformation** must be coded so as to make it replicable ($\approx$ reproducible),¹²⁷ which depends on the quality of code development and documentation (cf 5.b). This phase is all the more sensitive because several of the choices made during transformation cannot be reversed. Transformation in fully automated *pipeline* mode is the most practical solution: this is especially true when working with private data (Joubert 2025). In practice, however, we may find that certain steps in the transformation process require human intervention, resulting in multiple codes corresponding to successive automated steps.

¹²⁶ In Chapter 3 we looked at why the term “raw data” is problematic (Gitelman 2013), explaining that data considered to be “raw” by one user may appear highly constructed to another. The meaning of this adjective is unstable. Hence the importance of including the word “statistical” here, making clear that we are talking about the input to a statistical process, in much the same way that we might talk of “raw survey data”. Even with this caveat, we must admit that this is a semantically unstable solution.

¹²⁷ Two discrete concepts, in fact. Reproducibility is the capacity to precisely replicate the results of a study using the same data and the same code. This is a very high bar, best suited to scientific research. Replicability is the capacity to obtain similar results to those generated by the original study, using independent data and/or a new way of applying the same methods. This is closer to the spirit of statistics: we are not seeking to offer definitive proof. On this subject, see (National Academies 2019), cited above.

Finally, the results of **quality assessment** need not be recorded with the strictest degree of formality; they are primarily important for what they allow us to learn. They serve to reveal to us the reality of the data at the conclusion of Phase 2: undercoverage, dubious data, data with “missing values” (with clear rules on how these entries are to be identified). This information will be of great importance in the next phase of statistical processing, showing us where to focus our attention.

Figure 5 - Phase 2 - Transformation of administrative data



6. The statistical processing phase

Phase 3 is not fundamentally different from the processing operations applied to survey data, i.e. step 5 in the GSBPM, which is indeed entitled “Process”. As we have seen, the first two phases bring the data to a point where they are comparable to survey data, and from then onwards the process is the same.

6.a. Context

In this processing phase, we start with a “raw statistical” data file which is likely to possess certain defects: undercoverage, missing values, abnormal or dubious values, inconsistencies with other figures etc. We thus apply processing operations designed to eliminate or attenuate these defects (imputation, for example), then produce statistics, with feedback loops if needed, much as we would with survey data. Much work is required to fine-tune the data file, including a “cleaning” phase (Chu et al. 2016) which will be essential for future uses of the data.

Moving on from the preceding phases, one new element comes into play: we are now entirely free of the administrative context, having fully completed the transition into a statistical realm where we can count on an *infrastructure of knowledge* (Borgman 2015), incomplete though it may be.

In the context of official statistics, this infrastructure includes:

- A shared register of metadata (Bonnans 2019), including, for example, definitions of shared concepts or domains;
- Shared classification systems;
- Registers of individuals, households, homes (for Resil see Lefebvre 2024; Espinasse, Roux 2022), or establishments (Hachid, Leclair 2022; Bensoussan et al 2023);
- Widely recognised, documented and disseminated methods, and shared tools;
- Existing published statistics, allowing for comparisons;
- Seminars, colloquia, committees with the participation of experienced official statisticians etc.

The existence of this infrastructure makes things easier, providing a solid working environment for statisticians, but it does not solve all of our problems. One key consideration in Phase 3 is whether or not **we have a register of the population in which we are interested**. This has major implications for the processing phase as a whole.

If such a register exists, we can constantly refer back to it, as we do for surveys with the sampling frame. This guarantees a certain level of quality and allows for comparisons, particularly when it comes to estimating undercoverage, which was not feasible (or else only roughly) in Phases 1 and 2.

The Resil system (Lefebvre 2024), comprising administrative data, bases all of its operations on a “backbone” of registers: registers of individuals, homes and households which provide invaluable security guarantees. These registers can be cross-referenced with key sources constructed by official statisticians, such as Fideli (Lamarche, Lollivier 2021) for individuals, or Filosofi¹²⁸ for households, which offer points of comparison

128 Social and fiscal localized income : Fichier Localisé Social et Fiscal (Filosofi) system was set up to provide indicators of living standards, inequality and poverty at a local, sub-departmental level.
<https://www.insee.fr/en/metadonnees/source/serie/s1172>

With business statistics, the Sirius¹²⁹ register (Hachid, Leclair 2022) provides this backbone, cross-referenced with sources such as the DSN, giving rise to the “all-employee database” among other tools.

But such registers are not always available, and this has consequences for the rest of the process.

6.b. The processing operations

The standard operations are well known and have been thoroughly documented in statistical methodology papers over many years. This phase also includes operations designed to deal with item and unit nonresponse (Caron 2002), margin calibration (Deville, Särndal 1992), overcoverage or undercoverage (Wallgren, Wallgren 2007) and *data editing* (De Waal et al. 2011).

In addition to the working environment, Phase 3 has another distinctive characteristic: during this phase, we have the option of **modifying the data for statistical reasons**:

- Making use of models we consider to be relevant (for imputation in particular),
- But also by means of manual modifications performed by clerks.

This is not possible in Phase 1 (during which we do not alter the administrative data), nor in Phase 2 (where data are a product of transformation, not yet subjected to statistical analysis).

During processing, the usual statistical methods apply, and require certain adaptations in order to be implemented correctly. We will look in more detail at these adaptations hereunder, limiting ourselves to questions of scope, missing data and tabulation, and leaving data editing for the next section.

• Scope

We must distinguish clearly between statistical scope and administrative scope. In Phase 2, during the filtering phases, we eliminated much of the overcoverage problem with regard to statistical scope, but undercoverage of the administrative scope cannot be observed, because we have no points of comparison. Handling undercoverage is thus a task for Phase 3.

Generally speaking, it is possible to tackle problems of scope divergence¹³⁰ (statistical scope differing from administrative scope) and statistical undercoverage by means of imputations, additional data provided by other sources, or even weighting calculations, although the use of weighting and calibration when working with administrative data is still fairly rare in official statistical circles.

The way we go about this will depend on whether or not we have a register for our reference population.

129 SIRUS, the statistical business register, is linked to SIRENE and meets the requirements of the statistician: it takes into account the notion of enterprise in the economic sense, it integrates new units such as groups of companies, and it implements specific processing such as statistical cessation.

130 The problems are not the same as those observed in the preceding phase. In Phase 2, we were mostly concerned with questions of overcoverage, potentially requiring us to eliminate some information (or at least calculate a probability indicator). In this phase we will be primarily focusing on undercoverage, or problems of completeness, utilising other sources of information.

- *Scenario 1: a register exists for the reference population*

In this case, we need not worry about the distinction between statistical and administrative scope. We can observe the undercoverage of the statistical scope by looking at the difference between the “reference” population and the population actually present in the files derived from Phase 2. If necessary, we can complete the data by imputing total non-responses, or by using other sources where they exist.

- *Scenario 2: no register exists for the reference population*

Since we do not have the equivalent of a sampling frame which would allow us to determine the size of the population we are interested in, we need to turn to other data sources for information relevant to our scope, and to determine whether or not the units present and the expected/absent units do indeed fall within that scope. It all depends on the nature of the additional information at our disposal.

We may be aware of margins which allow us to perform calibration: it would thus become possible to deal with missing units by means of reweighting, but this approach needs to be deployed with caution.

Another possibility is that we might have alternative data sources which only partially cover the scope of our interest, but which nonetheless serve as benchmark sources. They might be used to evaluate the degree of coverage (much like the post-census surveys conducted to assess the coverage ratio of the census itself), or else be taken from existing data sources. Using control samples differs from the calibration method described above, and leads us to capture-recapture methods (Ardilly & Koumarianos 2025).

On a general level, working with multiple sources opens up new methods for evaluating the coverage ratio (multiple systems estimation, or capture-recapture (Chiperfield et al 2024)) or determining whether or not units fall within our scope, particularly thanks to the “signs of life” method (Óvári, Kočíš 2022, Tiit 2017).

Alternatively, we may have no information about the scope: in this case we can cross-compare information to deal with overcoverage.

For example, we can use information on deaths from the National Register for the Identification of Natural Persons to clean out the driving licence file. This, however, is an administrative use.

Another more complex example involving two difficulties (over- and undercoverage): determining the size of the resident population based on administrative data depends on the completeness of the sources, a quality to which many aspire but few achieve. In order to construct reliable data, statisticians can seek out positive or negative “signs of life”. Some countries use information such as school enrolment for children, non-renewal of driving licences, or use of medical services. Combining multiple sources of administrative information helps us to refine the probability that a unit belongs to the target population.

A frequently encountered problem is how to manage units that are absent from the administrative data: there is no equivalence between the presence of data pertaining to a unit, and the real existence of that unit. The demise of a unit in real life (e.g. a company which stops trading) may only be recorded later on in the administrative information system, whereas the absence of annual accounts

may provide an advance warning that the company is no longer in business. Units may be listed on old versions of administrative registers, and this does not mean that they are necessarily still active.

- ***Missing data***

In the preceding section on undercoverage, we naturally found ourselves confronted with the question of how to deal with missing data, in this case missing data units (we might compare them to “total non-responses” in surveys, although the absence of a sampling frame makes these two difficulties – non-responses versus missing units – very different in practice). In this section we focus primarily on missing data within existing units, the equivalent of partial non-responses, but some of what is said here is also applicable to units which are missing entirely.

In order to deal with this effectively, we first need to know more about the distribution of the missing data; this is one of the outcomes of the quality assessment conducted at the end of Phase 2.

On this subject, Little & Rubin (2002) describe three types of missing data distribution (univariate, uniform, unstructured). They formally describe the different mechanisms leading to missing data, which they divide into three categories: Missing Completely At Random (MCAR), Missing At Random (MAR), or Not Missing At Random (NMAR).¹³¹ This classification is not dependent on the way the data were obtained, and may thus be just as applicable to surveys as to administrative data (where MAR and NMAR will be the biggest problems). In the context of surveys, these mechanisms are connected to the *propensity to respond*... which is irrelevant with administrative data, since there are no questions to be answered.¹³²

With administrative data, the risk of bias induced by the absence of certain data, or statistical units, will need to be addressed: this time, understanding the origin of these absences (of units or data) requires us to re-examine the logic of administrative management processes. To do this, it may be useful to produce some exploratory statistics regarding the distribution of missing data across the population, using different variables.

We can then deal with missing data using imputation methods suited to the configuration at hand. Imbert and Vialaneix (2018) provide a good overview of the various configurations of missing data and corresponding methods, which are entirely applicable to administrative data. Of course, if we have another statistical source at our disposal which serves as a proxy for the data, then that is better still.

- ***Dubious data***

In Phase 2 we were able, with the help of a panoply of automated consistency/likelihood checks, to identify dubious data: these are not statistics which need to be handled with caution, but rather microdata. Phase 3, with its *data editing* operations, also produces this kind of information.

Nevertheless, data which appear to be abnormal may turn out to be entirely accurate.

For example, a business may report very high turnover despite having virtually no employees. This often happens when companies rely heavily on agency workers.

How do we deal with these dubious data? In each case, we have a choice of options:

¹³¹ A correlation between the absence of information and variables of interest (NMAR) may lead to biased estimates, irrespective of the origin of the data (surveys or administrative data).

¹³² With the potential exception of administrative declarations, and even then...

- Leave the value unchanged,
- Replace it with an imputed value,
- Replace it with the corresponding value from another source, where available.

The choice to leave the value as it is may be made on the basis of automated criteria, establishing a sort of threshold for inconsistency above which data are discarded. Or else, if the number of cases is too great, we may ask data clerks to perform manual checks and decide whether or not the data should be imputed. The big difference compared with surveys is that we cannot check with the original provider of the data.

As for imputation itself, the situation is similar to that encountered with missing data, and the process is no different from that applied to survey data. We can use data correction techniques to ensure that the relevant constraints are satisfied: these methods are derived from Fellegi & Holt's seminal article (1976).¹³³

- **Tabulation**

This is the final processing operation: we are no longer concerned with refining a file containing individual data; we are now using that file to calculate statistical aggregates. Tabulation with administrative data is no different from tabulation with survey data, apart from the fact that we have no survey weight, and if we do not have a reference population then the risk of error can be very substantial. The feedback loop thus has a major role to play here.

6.c. Quality assessment and feedback: the importance of *output editing*

In Phase 3, the main priority of quality assessment is to verify the plausibility of the statistical aggregates. Technically speaking, the process is fairly similar to that seen in Phase 2, seeking to verify: consistency with known totals, consistency in the variations of the totals, completeness of scope and plausibility of a few distributions. But while Phase 2 was essentially about conducting elementary statistical checks to ensure that there were no errors of scope, in Phase 3 our aim is to check the statistics (aggregates, distributions) against our variables of interest, for validation purposes.

These are final checks conducted on the statistics (not the microdata): to draw a parallel with *data editing* as applied to survey data, this corresponds to the part we call *output editing* (Arbués et al 2013). To begin with, it consists of exploring the statistics available for similar subjects, observing their variations to ensure that the orders of magnitude are correct. If inconsistencies remain, we must work out why: the feedback loop will need to track back to the preceding phases, to Phase 2 or even Phase 1, often to deal with misunderstandings relating to the administrative domain. But things are nowhere near as simple as they are when dealing with survey data, because the statistics are no longer expressed in the language of the administrative entity: statisticians must thus “make do” with the elements at their disposal.

Considering the rounds of quality assessment, conducted in Phases 1, 2 and 3 respectively, it seems clear that *output editing* remains the ultimate and most important verification. We might envisage lightening the verification load in the preceding steps, but output editing is non-negotiable. It provides the safety net which official statisticians need before they can publish anything.

¹³³ After almost 50 years, this method remains a benchmark when it comes to performing corrective measures *to verify the control rules*. The Fellegi-Holt method minimises the number of corrective measures required to achieve this goal.

It is true that the quality assessment conducted at the end of Phase 2 has much in common with the operation conducted in Phase 3. The advantage of performing quality assessment in Phase 2 is that it allows us to confirm the success of the transformation process. But we must admit, if any of these three rounds of quality assessment were to be sacrificed, it would have to be number 2.

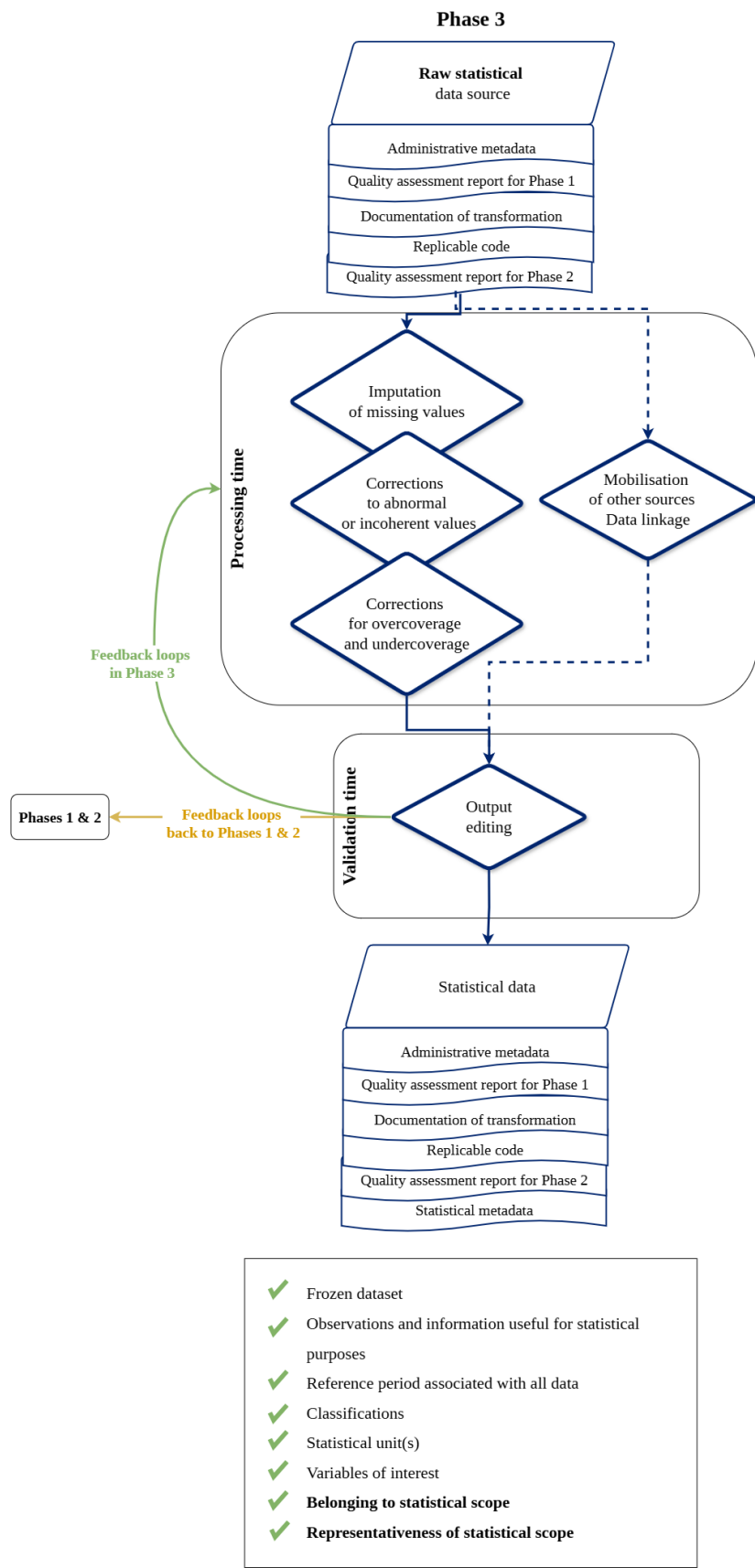
6.d. Outcomes of the statistical processing phase

At the conclusion of this statistical processing phase, we thus find ourselves with:

- A completed statistical dataset in which missing values have been handled, and which is consistent with the statistical scope
- Statistical tables
- The results of the quality assessments.

Generally speaking, it is the third of these outcomes which is the least dependable: we do not always keep proper records of the checks we perform, despite the fact that this information can be useful if the source needs to be evaluated multiple times, or if there are several types of users.

Figure 6 - Phase 3 - Processing statistical data



C. Conclusion

7. Summary and outlook

In January 2024, Peter Christen and Rainer Schnell published a short article examining the many forms of misunderstanding which continue to shroud data acquisition, processing and data linkage. They even presented a list of false preconceptions for each of these operations. Perusing those pages, one can sense the authors' irritation at the stubborn errors committed by data scientists, led astray by an overly simplistic vision of what data truly are. Although the article addresses data science in general, the authors' observations are wholly applicable to statistics.

The increasingly systematic use of new data sources appears to be inevitable for official statistics: the *datafication* of the world (Rey 2016) and the multiplicity of data sources, ever-increasing data-access, storage and computational capacities, ever-increasing powers of calculation and analysis... there is every reason to believe that statistics will become even more *data-driven* in the future (Salemink et al. 2020). But, as Christen and Schnell rightly note, the challenges involved are many and various.

We have deliberately limited the scope of this working paper: the “exclusive” use of administrative data (to borrow Rancourt's (2018) terminology), focusing on the method employed to produce statistics, including quality assessment (Laitila et al. 2011; Eurostat 2016a), but omitting indirect uses and the challenges of multi-source environments.¹³⁴ In doing so we have had to confront, much like Christen and Schnell, the fundamental flaws inherent to these data when viewed from a statistical perspective. What we need now are dedicated mechanisms for acquiring administrative data for statistical purposes (FAO 2018), systems capable of handling this endemic imperfection, as well as the distance between the producers and users of administrative data (Borgman et al. 2025)... users who are themselves producers of statistical data.

7.a. Summary of our proposed approach

With a view to creating such mechanisms, this working paper proposes a general approach in the form of an analytical framework, accompanied by a three-phase process, with each phase presenting its own challenges.

We propose a conceptual vision of the problem, which could be summarised as follows.

Phase 1 can be seen as an exercise in projection: we start in an “administrative” space with many dimensions (the information system), one of these dimensions being time. The aim of this phase is to project the data into a smaller “administrative” sub-space (the administrative source), by freezing the time dimension. The purpose of the operation is thus to embrace the immense range of possibilities but focus only on those which we consider to be necessary, all while harmonising the data and establishing fixed temporal reference points.

However, this image alone will not suffice: the aim is also, and above all, to take into account our lack of knowledge concerning the entire ecosystem underlying the administrative data. In truth, the main objective of the first phase is to manage *data friction*, as defined by Edwards et al. (2011). This requires a capacity to adapt to a “foreign” culture, immersing ourselves in this world in order to gain a detailed understanding of the ins and outs of the administrative processes, before gradually establishing a formal agreement with the entity providing the data. As things progress it soon

¹³⁴De Waal et al. (Eurostat 2016 b) list 6 uses: two direct uses (direct tabulation, and use as either a substitute or supplement for another data source) and four indirect uses (creation and upkeep of registers, *data editing* and imputation, indirect estimates, and data validation). Irrespective of the proposed use, statisticians always find themselves faced with various integration challenges.

becomes clear that no aspect of this exercise is self-evident, and that we must shake off the naïve, simplistic temptation to say: “just send us the data, we’ll work it out.”

Phase 1 is thus largely devoted to preparation, laying the groundwork for the actual process of receiving and acquiring the data. An analytical framework is deployed to facilitate this process, because data do not just exist “in the ether”: they abide by a certain grammar, defining their scope, their objects, their variables, their domains and their time frames.

To continue with the mathematical analogy, in some respects **Phase 2 is a kind of morphism**: we start with an administrative sub-space E , with a clearly defined reference date/period and a reduced number of dimensions. We then transport it into another space, along with the structure associated with the data (object, scope, variable, domain, time frame). In spirit, this is akin to constructing an f function so that $F = f(E)$, a function which is also capable of transporting the analytical framework.

A theme which runs throughout this 2nd phase is the realisation that we cannot work directly with administrative data in their original state, as their grammar and semantics are too far removed from the analytical requirements of official statistics. It is therefore essential that, once the data have been acquired and their quality evaluated, we extract them from the administrative context (detachment) so that we repackage and transform them into a form solidly compatible with the statistical realm. Here again, there is an analytical framework suited to this specific purpose. The transformation, i.e. our f function, is an object in its own right, which must be documented, maintained and updated.

This is an intermediate, hybrid phase between the administrative and statistical realms, a transitional phase which allows us to move on to the next step.

Phase 3 is where the real statistical work happens, the phase in which professional statisticians feel most at home. Here they are able to operate in an entirely statistical environment, with its familiar codes, methods and lexicon.

This is the first phase in which it becomes possible to modify the data, either manually or using models, to ensure that they fulfil some criteria (such as completeness, or the absence of missing values). We can also deploy a whole array of familiar statistical operations on the microdata, all the way up to tabulation, i.e. the production of actual statistics, which marks the culmination of the process.

In the 3rd phase, we no longer have to worry about data friction ... well, almost. This is also an opportunity to spot inconsistencies, and trace their causes back to earlier phases in the process.

Which serves to illustrate another aspect of our proposed approach: the presence of “**feedback loops**” in each phase. There is nothing particularly theoretical or abstract about this: in the acquisition phase, for example, it simply means that if the statisticians are not satisfied with the data provided, they should be capable of identifying the reasons for this dissatisfaction by means of automated quality assessment, justifying a request for a new delivery of data or metadata, or simply additional explanations from the party providing the data. In Phase 2 (transformation), feedback is more focused on the transformation pipeline, while in Phase 3 we have a more traditional *data editing* process, with final iterations and backtracking where necessary.

These 3 phases are inevitably **interconnected**, not least by means of the feedback loops, but each phase yields data in different “states”: still in the administrative realm in Phase 1, moving into the statistical realm for Phases 2 and 3.

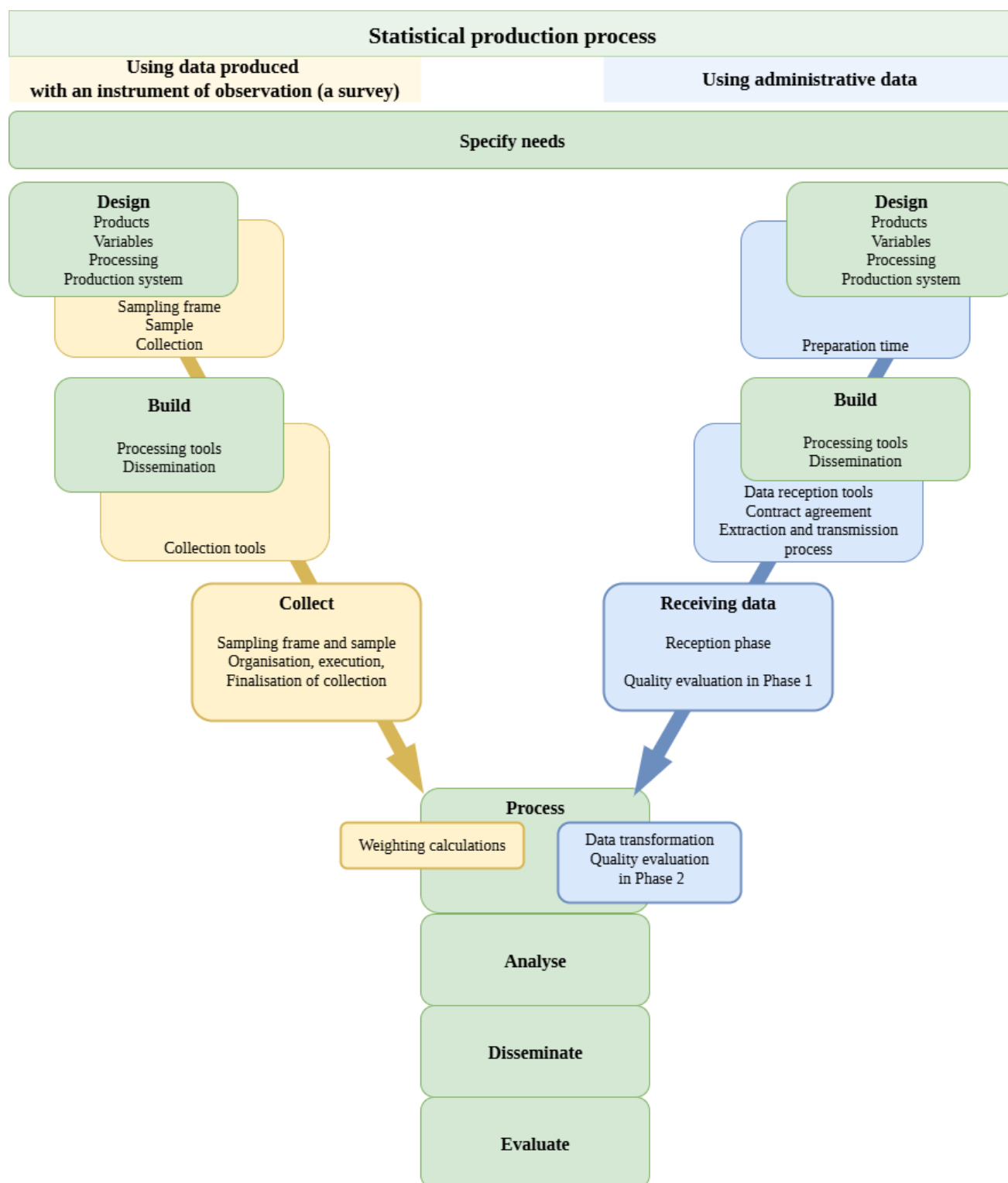
Each phase advances the process and addresses a distinct set of problems, making contributions which are summarised hereunder:

Figure 7 – Transition from the administrative analytical framework to the statistical analytical framework

Existing information in the administrative realm	→	What we need in the statistical realm	Phase 1	Phase 2	Phase 3
live information system		frozen dataset			
very rich information system		observations and information useful for statistical purposes			
no collection period		reference period associated with data			
no reference period					
Lists of administrative codes		Classifications			
Administrative objects		Statistical unit(s)			
Administrative concepts		Variables of interest			
Real administrative scope		Belonging to statistical scope			
		Representativeness of statistical scope			

The 3 phases also allow us to complete the Y diagram introduced at the end of Chapter 2, where we deliberately left the branch for administrative data empty. The completed diagram is shown overleaf.

The complete Y diagram



This diagram is a completed version of that shown on page 24, this time including the sub-phases connected with the use of data from administrative sources.

The Process phase comprises a number of steps described in Chapters 5 and 6: some involve certain specificities when working with data from administrative sources (for example, quality assessment at the end of Phase 2), whereas others are similar (calculating aggregates etc.). Full details of the sub-phases can be found in the GSBPM (UNECE, 2019).

7.b. Implementation

This framework may appear overly theoretical or idealised. In practice, can the phases always be separated so clearly?

What may prove to be impossible, in certain cases, is to establish a feedback loop involving the administrative data provider. It may turn out that the latter is not interested in any additional exchanges, or simply does not have the capacity to deal with statistical requests. Administrative bodies have their own operational objectives, and to them the production of statistics amounts to a third party making secondary use of existing data.

Implementing a three-phase approach, with a clear separation between each phase, may also prove to be difficult on an operational level: difficulties storing very large volumes of data, handling the workload etc. In practice, this often leads statisticians to merge the phases. Phases 1, 2 and 3 provide a framework and, practically speaking, the **important thing is to know why we might decide to stray from the framework, and what the consequences could be**. To begin with, let us attempt to analyse the impact of failing to separate the acquisition phase from the transformation phase, or the transformation phase from the processing phase.

We must first make a distinction between two possible scenarios: one where we are acquiring data for the first time, and another where the same administrative source has been used multiple times. Maintaining a clear separation between the phases is more important in the first of these scenarios.

Acquiring data for the first time

The preparatory work which makes up Phase 1 takes various guises, and cannot be skipped: acclimatising to the administrative context, documentation, metadata, initial contractual arrangements. The separation between quality assessment in Phase 1 (with accompanying feedback loop) and Phase 2 is essential: if we do not evaluate quality at this juncture, problems are likely to arise at a later stage. When we find ourselves with transformed data, it will be much more difficult to interact with the administrative entity, and thus to ask the data provider to come up with some better-quality files.¹³⁵ Moreover, if quality assessment is delayed until Phase 2, it becomes harder to untangle problems found in the source from problems induced during transformation. In general, quality assessment is very important, and special attention must be devoted to missing units (objects) and variables, as well as abnormal variables. Working with the new GMBI source, for example, comparison with data from the land register has proved to be indispensable.

Considering Phase 2 in isolation, we might imagine that it could be merged with Phase 3. However, doing so leaves us with a single *pipeline* containing both "non-statistical" (transformation)¹³⁶ and statistical processing operations. This means that there are no fundamental checks which might allow us to trace problems back and resolve them, and no "freeze frame" giving us a clear understanding of what we have at our disposal before proceeding with the statistical operations.

Regular data acquisition

In Phase 1, from the second round of data acquisition onwards, we might assume that the preparation work was already taken care of. And yet, the context may have changed in the meantime: new variables, new domains, a new scope. These changes can be of real importance. Certainly, after the first round of acquisition we may be better acquainted with the culture of the

¹³⁵ At a push, this might not be too much of a problem when the "distance" between the administrative and statistical realms is limited.

¹³⁶ Non-statistical in the sense that these operations are not traditional statistical methods, and transformation could apply to other domains.

administrative partner providing the data, but we still need to fine-tune or even amend the data sharing agreement, managing changes in the data and metadata using the appropriate framework (scope / variable / domain). Ideally, when acquiring data for the first time, the data receipt and quality-assessment phase will be well documented, with useful tools developed. The importance of updating these documents and tools must not be overlooked, and the workload is likely to be less substantial than it was first time around.

We might thus imagine integrating Phase 2 into Phase 1, receiving data which are quality-assessed then transformed, or transformed and then quality-assessed. But caution is required, because we must always retain the option of exchanging feedback with the administrative entity that provided the data, and this becomes a much more complex proposition once the data have been transformed. Moreover, during quality assessment it should be a little less complicated to distinguish between problems connected with the source and problems arising from transformation. It is important that we analyse the volume of data received, before beginning transformation.

If we were to merge Phase 1 and Phase 2 we would end up with a broad data receipt – quality assessment – transformation phase, followed by a statistical processing phase.

Another option is to fold Phase 2 into Phase 3, which amounts to postponing all statistical quality assessment until the *output editing* step. Nonetheless, quality assessment is essential when things change: a change of scope from one time period to the next can become a major problem, and the most likely source of risk.

If we merge Phases 2 and 3, we thus end up with a data receipt – quality assessment phase, followed by a transformation – processing phase which is as automated as possible. The most important thing is to be capable of spotting the most consequential errors, and putting a feedback loop in place. As a general rule, if we decide to minimise the number of quality assessment operations then we run the risk of putting it all off until the final step, and the essential process of *output editing* thus becomes the sole remaining “safety net”, albeit with limited power to intervene and course-correct.

Pooled data acquisition

Looking beyond the initial context (an administrative source,¹³⁷ and statistics based on that source), the first subject which springs to mind is the possibility of **pooling**: for a given data source, if we have established a clear and rigorous framework for a specific use, it would be inefficient not to reuse at least some of it. At the very least we can share documentation, metadata and knowledge of the subject matter. Depending on the use, pooling may occur at different levels (at the end of Phase 1, Phase 2 or Phase 3), or encourage us to make Phase 2 as open as possible in its design. This design must also be well documented, so that we can pass on to other potential users a solid understanding of the transformations performed.

7.c. Further developments

Thoroughly understanding a source opens up the possibility of using it in a multi-source capacity, but for simplicity’s sake this paper has focused exclusively on single-source uses. It is, of course, common to work with multiple sources, both administrative data sources and surveys, within the same process. Multi-source statistics present other challenges not explored here. For De Waal et al. (2020), data integration issues can often be boiled down to three principal difficulties: harmonisation of units and concepts, selection of relevant values, and data linkage. We also need to consider estimation methods making simultaneous use of different sources (composite estimates).

¹³⁷ Or a coherent set comprising several administrative sources (cf. land register data, for example).

The issue of accuracy (mean squared error, variance) has not been addressed at all, and this is a subject which we can ill afford to neglect. Relatively little research has been conducted on the estimation of total survey error outside the context of surveys, but some useful contributions are available (Groves, Lyberg 2010). It requires work to model the different types of error, without the central presence of sampling error.

Above and beyond these abstract questions of calculation methods and tools, there is no way of using administrative data without considering questions of confidentiality, ethics and acceptability: these data should not be treated as mere inputs to be plugged into an industrialised process. They are particularly sensitive because they relate to individuals, who may not wish to see them re-used, and who have not necessarily given their approval, which is of course very different from the way things work with surveys.

Finally, another possible avenue for building upon this research would be to apply it to private data. Numerous examples have been cited throughout this document, and in truth there is not all that much difference between producing statistics using administrative data and using private data. There are some differences, of course. Lesur (2025) mentions problems including completeness, representativeness, the risk of duplicate counting (between operators, for example), the increased complexity of the legal framework, and the fact that the data are not free.¹³⁸ Since it is generally difficult, or impossible, to directly access information systems, *pipelines* are often put in place to handle the transformations and perhaps even the first round of aggregation: statisticians are thus only able to access the results which emerge from the *pipeline* (Joubert 2025). This is akin to combining Phase 1 and Phase 2, and the prospect of establishing a feedback loop with the data provider is far from straightforward.

¹³⁸ Since 2016, Law No. 2016-1321 dated 7 October 2016, known as the Digital Republic Act, has theoretically made data sharing compulsory and free of charge between administrative entities.

Bibliography

AMOSSE, Thomas, 2020. The 2020 Socio-Professional Classification: Continuity and Innovation, For Greater Use. *Courrier des statistiques* N° N4, juin/june 2020.

ANXIONNAZ, Isabelle et MAUREL, Françoise, 2021. The National Council for Statistical Information (Conseil national de l'information statistique): the quality of Official Statistics also depends on consultation. *Courrier des statistiques* N° N6, july 2021.

ANDRE Mathias, ARNOLD, Céline, et MESLIN Olivier, 2021. 24 % des ménages détiennent 68 % des logements possédés par des particuliers. France Portrait Social 2021, Insee.

ANDRE Mathias et MESLIN Olivier, 2022. Property wealth of households - Lessons from the use of comprehensive administrative data sources combination. In : *Courrier des statistiques* N° N7, january 2022.

ARBUÉS, Ignacio, REVILLA, Pedro, et SALGADO, David, 2013. An Optimization Approach to Selective Editing. *Journal of official statistics* Vol. 29, No. 4, 2013, pp. 489–510.

ARDILLY, Pascal, 2006. Les techniques de sondage. Nouvelle éd. actualisée et augmentée. Paris : Éd. Technip. ISBN 978-2-7108-0847-3.

ARDILLY, Pascal, KOUMARIANOS, Heidi, 2025. Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques. Insee, Document de travail N° M2025-01, Janvier 2025.

d'AURIZIO Leandro et PAPADIA Giuseppina, 2019. Using Administrative Data to Evaluate Sampling Bias in a Business Panel Survey. *Journal of Official Statistics*, Vol. 35, No. 1, 2019, pp. 67–92.

BABB, Penny, 2017. The assurance of administrative data: A proportionate approach. *Statistical Journal of the IAOS* 33 (2017) 435–440. DOI 10.3233/SJI-160321

BACH, Laurent, BOZIO, Antoine, DUTRONC-POSTEL, Paul, FIZE, Étienne, GUILLOUZOUIC, Arthur et MALGOUYRES, Clément, 2023. Évaluation de la réforme de la taxe d'habitation. Rapport IPP n°48 – Décembre 2023

BAKKER B., van ROOIJEN J., et van TOOR L., 2014. The System of social statistical datasets of Statistics Netherlands : An integral approach to the production of register-based social statistics, *Statistical Journal of the IAOS* 30 (2014) 411–424

BATES, J., 2017. The politics of data friction. *Journal of Documentation*. ISSN 0022-0418

BATINI, Carlo et SCANNAPIECO, Monica, 2016. Data and Information Quality – Dimensions, Principles and Techniques. *Springer*. ISBN 978-3-319-24104-3.

BECK, François, CASTELL, Laura, LEGLEYE, Stéphane et SCHREIBER, Amandine, 2022. Mixed-mode collection in household surveys: a modernised collection method and a more complicated process, *Courrier des statistiques* N° N7, january 2022.

- BENICHOUE, Yves-Laurent, ESPINASSE, Lionel, et GILLES, Séverine, 2023. The Non-Significant Statistical Code (CSNS): a service to facilitate file matching, *Courrier des statistiques* N° N9, juin 2023.
- BENSOUSSAN, Johanna, BIZINGRE, Joël, et COURVALIN, Nathalie, 2023. FINESS, the register of healthcare establishments, *Courrier des statistiques* N° N10, december 2023.
- BERTHE, Juliette, 2025. Le défi des données pour l'inspection générale des affaires sociales, *Courrier des statistiques* N° N13, juin 2025
- BERTHELOT, Gregory, 2020. Premier bilan du prélèvement à la source après un semestre de mise en œuvre : une entrée en réforme réussie, *Revue française de finances publiques* 2020/1 N°149.
- BERZOFSKY Marcus E., LIAO Dan, BARNETT-RYAN Cynthia, SMITH Erica L., 2025. A total data quality paradigm for official statistics based on administrative data. *Statistical Journal of the IAOS*. 2025;. doi:[10.1177/18747655251337671](https://doi.org/10.1177/18747655251337671)
- BIEMER, Paul P., de LEEUW, Edith, ECKMAN, Stephanie, EDWARDS, Brad, KREUTER, Frauke, LYBERG, Lars E., TUCKER, N. Clyde, WEST, Brady T. (Eds), 2017. Total survey error in practice. Wiley, 2017.
- BOLLIET, Quentin, FLOYRAC, Aymeric, MAILLARD, Sophie, et ROSENZWEIG, Agathe, 2025. L'inspection générale des finances et la science des données - Quelles méthodes pour quels usages ?, *Courrier des statistiques* N° N13, juin 2025
- BONNANS, Dominique, 2019. RMéS: INSEE's Statistical Metadata Repository, *Courrier des statistiques* N° N2, 27 june 2019.
- BONNET Odran, et LOISEL Tristan, 2024. L'économie racontée par les données bancaires, *Courrier des statistiques* N° N12, décembre 2024
- BONTEMS Thierry, et GOULIN, Sabine, 2010. De la Qualité de l'Information Administrative et Scientifique. 8. *Ecole Inter-organismes "Qualité en Recherche et en Enseignement Supérieur"*, Montpellier, Sep 2010.
- BORGMAN, Christine L., 2015. Big data, little data, no data : scholarship in the networked world. *The MIT Press*, 2015.
- BORGMAN, C. L., DARCH, P. T., SANDS, A. E., et GOLSHAN, M. S. , 2016. The Durability and Fragility of Knowledge Infrastructures: Lessons Learned from Astronomy. In *Proceedings of the 79th ASIS&T Annual Meeting* (Vol. 53). Copenhagen: ASIS&T. <https://www.asist.org/files/meetings/am16/proceedings/submissions/papers/31paper.pdf>
- BORGMAN, Christine L., GROTH Paul T., 2025. From data creator to data reuser : distance matters. *Harvard Data science Review*.
- BOWKER, Geoffrey C. et STAR, Susan Leigh, 2000. Sorting things out. Classification and its consequences. 25 août 2000. *The MIT Press*. ISBN 978-0262522953
- BOYDENS, Isabelle, 1999. Informatique, normes et temps – Évaluer et améliorer la qualité de l'information : les enseignements d'une approche herméneutique appliquée à la base de données «LATG» de l'O.N.S.S. *Éditions E. Bruylant*. ISBN 2-8027-1268-3.

- BOYDENS, Isabelle, 2018. Data Quality & « back tracking » : depuis les premières expérimentations à la parution d'un Arrêté Royal, *Smals Research*, 2018 <https://www.smalsresearch.be/data-quality-back-tracking-depuis-les-premieres-experimentations-a-la-parution-dun-arrete-royal/>
- BOYDENS, Isabelle, 2021. Data Quality Tools : retours d'expérience et nouveautés. *Smals Research*, 2021 <https://www.smalsresearch.be/data-quality-tools-retours-dexperience-et-nouveautes/>
- BOYDENS Isabelle, HAMITI Gani G., et VAN EECKHOUT Rudy, 2021, A service at the heart of database quality: presentation of an ATMS prototype , *Courrier des statistiques* N° N6, july2021.
- BRACKSTONE, G.J. 1987. Issues in the Use of Administrative Records for Statistical Purposes. *Survey Methodology*, vol. 13, pp. 29 – 43
- BRION Philippe, Esane, le dispositif rénové de production des statistiques structurelles d'entreprises. *Courrier des statistiques* n° 130, mai 2011
- BRUMBERG H.L., DOZOR D.,et GOLOMBEK S.G., 2012. History of the birth certificate: from inception to the future of electronic data. *Journal of Perinatology* (2012) 32,407–411
- De BROE Sofie, STRUIJS Peter, DAAS Piet, Van DELDEN Arnout, BURGER Joep, Van Den BRAKEL Jan, Ten BOSCH Olav, ZEELLENBERG Kees, et YPMA Winfried, 2020. Updating the Paradigm of Official Statistics : New Quality Criteria for Integrating New Data and Methods in Official Statistics, *Working paper* 02-20, *Center for Big Data statistics*, CBS
- BYCROFT C., et MATESON-DUNNING N., 2020. Use of administrative records for non-response in the New Zealand 2018 Census. *Journal of the IAOS*, 36 (2020) 107–116
- CAMUS, Benjamin, 2022. The challenge of developing a statistical classification of crimes. *Courrier des statistiques* N° N7, january 2022.
- CARON, Nathalie, 2002. la correction de la non-réponse par repondération et par imputation. Document de travail M0502, Insee, Série des documents de travail « Méthodologie statistique ».
- CARON, Daniel J., NICOLINI, Vincent, et BERNARDI, Sara, 2020. Réflexion sur les stratégies de données gouvernementales : rapport de recherche. Ecole Nationale d'Administration Publique, Chaire de recherche en exploitation des ressources informationnelles, Bibliothèque et Archives nationales du Québec, mars 2020.
- CERRONI, Fulvia, BELLA, Grazia Di et GALIE, Lorena, 2014. Evaluating administrative data quality as input of the statistical production process. *Rivista di statistica ufficiale*. 2014. N° 1, pp. 30.
- CHALEIX, Mylène, et VANDERSCHULDEN, Mélanie, 2023. Lignes directrices sur la coopération entre acteurs pour l'exploitation statistique de données administratives, *Document interne Insee*, septembre 2023
- CHALEIX, Mylène, et MIKOL, Fanny, 2024. Accès du service statistique public aux données privées : les étapes-clé, les acteurs, les points de vigilance, *Document interne Insee*, septembre 2024
- CHAMBAZ, Christine, 2018. De l'activité de la justice au suivi du justiciable - Faire parler les données de gestion. *Courrier des statistiques* N° N1, décembre 2018.

- CHIPERFIELD, James, CHU, Randall, ZHANG Li-Chun, et BAFFOUR Bernard, 2024. Robust Statistical Estimation for Capture-Recapture Using Administrative Data. *Journal of Official Statistics* 2024, Vol. 40(2) 215–237
- CHRISTEN, Peter et SCHNELL, Rainer, 2024. When Data Science Goes Wrong: How Misconceptions About Data Capture and Processing Causes Wrong Conclusions. *Harvard Data Science Review*, 31 janvier 2024. Vol. 6, n° 1. DOI 10.1162/99608f92.34f8e75b.
- CHU, Xu, ILYAS, Ihab, KRISHNAN, Sanjay, WANG Jiannan, 2016. Data cleaning : overview and emerging challenges. SIGMOD '16: Proceedings of the 2016 International Conference on Management of Data, pp. 2201 – 2206, <https://doi.org/10.1145/2882903.29125>
- COTTON Franck, et HAAG Olivier, 2023. Integrating administrative data into a statistical process - Industrialising a key phase. *Courrier des statistiques* N° N9, june 2023.
- COUDIN, Elise, et ROBERT, Aude, 2024. Les statistiques sur les causes de décès - Classer et coder... dans la classification internationale des maladies. *Courrier des statistiques* N° N12, décembre 2024.
- COURMONT Antoine, 2021. Quand la donnée arrive en ville. Open data et gouvernance urbaine. *Presses univ. de Grenoble*, « *Libres cours Politique* », ISBN : 9782706147357.
URL : <https://www.cairn.info/quand-la-donnee-arrive-en-ville--9782706147357.htm>
- DAAS, Piet, 2009. Checklist for the Quality evaluation of administrative data sources. *Discussion paper Statistics netherlands*. 2009. pp. 36.
- De WAAL T., PANNEKOEK, et J., SCHOLTUS S., 2011, Handbook of statistical data editing and imputation, Wiley.
- De WAAL, Ton, Van DELDEN, Arnout et SCHOLTUS, Sander, 2020. Multi-source Statistics: Basic Situations and Methods. *International Statistical Review* (2020), 88, 1, 203–228
- DENIS, Jérôme, 2018. Le travail invisible des données – Éléments pour une sociologie des infrastructures scripturales. Août 2018. *Presses des Mines, Collection Sciences Sociales*.
- DENIS, Jérôme, et GOETA, Samuel, 2017. « La fabrique des données brutes ». Ouvrir, partager, réutiliser, édité par Clément Mabi et al., *Éditions de la Maison des sciences de l'homme*, 2017, <https://doi.org/10.4000/books.editionsmsmh.9050>.
- DESFORGES, Corinne, 2021. Réforme de l'état et gestion publique. *Revue française d'administration publique* n° 180, 2021, p. 1105-1158.
- DESROSIERES Alain, 1993. La politique des grands nombres : histoire de la raison statistique. *La découverte*.
- DESROSIERES Alain, 2004. Enquêtes versus registres administratifs : réflexions sur la dualité des sources statistiques , *Courrier des statistiques* N° 111, septembre 2004.
- DEVILLE, Jean-Claude, 1997. Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?, Paris : Insee, 1997, *Document de travail INSEE. Unité méthodologie statistique*, 9701, 8 p

- DEVILLE, Jean-Claude, et SÄRNDAL, Carl-Erik, 1992. Calibration estimators in survey sampling. *Journal of the American statistical Association* 87 (418), 376-382
- DI RUOCCO, Nunzio, SCHEIWILER, Jean-Michel et SOTNYKOVA, Anastasiya, 2012. La qualité des données : concepts de base et techniques d'amélioration. In : BERTI-ÉQUILLE, Laure, 2012. La qualité et la gouvernance des données au service de la performance des entreprises. *Hermes Science Publications*. pp. 25-54. ISBN 978-2-7462-2510-7.
- DUBRULLE, Bertrand, ROSEC, Olivier, SUREAU, Christian, 2023. An exchange standard to feed repositories and to guarantee quality . *Courrier des statistiques* N° N9, june 2023.
- DUPONT Françoise, DUSSART Josy, GUILLAUMAT-TAILLIET François. The Consultation Process: an Essential Step for the RÉSIL Project. *Courrier des statistiques* N° N11, june 2024.
- EDWARDS, Paul N., 2010. A vast machine: computer models, climate data, and the politics of global warming. First paperback edition. Cambridge, Massachusetts London, England : The MIT Press. Infrastructures series. ISBN 978-0-262-51863-5.
- EDWARDS, P., MAYERNICK, M., BATCHELLER, A., BOWKER, G., et BORGMAN, C., 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5), 667-690.
- ELBAUM Mireille, 2018. Les enjeux des nouvelles sources de données , *Chroniques*, n° 16, CNIS, septembre 2018.
- ELMASRI, Ramez, et NAVATHE, Shamkant, 2016. Fundamentals of Database Systems, *Global Edition, Pearson*.
- ERIKSON, Johan, 2020. Using a Process Model at Statistics Sweden: Implementation, Experiences and Lessons Learned. *Courrier des statistiques* N° N4, june 2020.
- ESAYAS, Samson Yoseph, 2015. The role of anonymisation and pseudonymisation under the EU data privacy rules, *European Journal of Law and Technology* Vol 6, No 2 (2015).
- ESPINASSE, Lionel, ROUX, Valérie, 2022. The National Register for the Identification of Individuals (RNIPP) at the core of French administrative life. *Courrier des statistiques* N° N8, Insee, november 2022.
- EUROSTAT, 2016a. Methodology for data validation 1.0. Essnet Validat Foundation, juin 2016.
- EUROSTAT, 2016b. Estimation methods for the integration of administrative sources. Deliverable D1. Décembre 2016 (version 3).
- EUROSTAT, 2017a. European Statistics Code of Practice. 2017.
- EUROSTAT, 2017b. Good practices in accessing, using and contributing to the management of administrative data. Eurostat. WP1
- EUROSTAT, 2017c. Estimation methods for the integration of administrative sources. Deliverable D2. Janvier 2017 (version 3).
- EUROSTAT, 2021. European Statistical System Handbook for quality and metadata reports. *Eurostat manuals and guidelines*, 2021, re-edition.

FAO, 2018. Guidelines on improving and using administrative data in agricultural statistics. Food and Agriculture Organization of the United Nations, Février 2018.

FELLEGI, I.P., HOLT, D., 1976. A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, Vol. 71, No. 353 (Mar., 1976), pp. 17-35

FERMOR-DUNMAN, Verena et PARSONS, Laura, 2022. Data Acquisition processes improving quality of microdata at the Office for National Statistics. *Q2022*, Vilnius.

FOLEY Brian, SHUTTLEWORTH, Ian, and MARTIN, David, 2018. Administrative Data Quality: Investigating Record-Level Address Accuracy in the Northern Ireland Health Register. *Journal of Official Statistics*, Vol. 34, No. 1, 2018, pp. 55–81, <http://dx.doi.org/10.1515/JOS-2018-0004>

FRIEDL, Jeffrey E.F., 2006. Mastering regular expressions - Third Edition, O'Reilly.

FROESCHL, Karl A. et GROSSMANN, Wilfried, 2000. The Role of Metadata in Using Administrative Sources. *Research in Official Statistics*, 2000.

GARTNER, Richard, 2016. Metadata – Shaping knowledge from Antiquity to the Semantic Web. Springer.

GIRARD-CHANUDET, Camille, 2023. Le travail de l'Intelligence Artificielle : concevoir et entraîner un outil de pseudonymisation automatique à la Cour de Cassation - *RESET (Recherches en sciences sociales sur Internet)*, Technologies numériques et apprentissage.

GITELMAN, Lisa (éd.), 2013. « Raw data » is an oxymoron. Cambridge, Mass. : MIT Press. *Infrastructures series*. ISBN 978-0-262-31232-5.

GOETA, Samuel, 2024. Les données de la démocratie. *C&F éditions, Collection Société numérique*, février 2024.

GRANQUIST, Leopold, 1997. The new view on editing. *International Statistical Review* (1997), 65, 3, 381-387.

GROVES, Robert, et LYBERG, Lars, 2010. Total survey error – Past, present and future. *Public Opinion Quarterly*, Vol. 74, No. 5, 2010, pp. 849–879.

GUIBERT, Bernard, LAGANIER, Jean et VOLLE, Michel, 1971. Essai sur les nomenclatures industrielles. *Économie et statistique*. Février 1971. Insee. N° 20, pp. 23-36.

HACHID, Ali, et LECLAIR, Marie, 2022. Sirius, the business register for statisticians. *Courrier des statistiques* N° N8, november 2022.

HAND, David J., 2018. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* [en ligne]. juin 2018. Vol. 181, n° 3, pp. 555-605. DOI 10.1111/rssa.12315.

HEDLIN, Dan, 2003. Score functions to reduce business survey editing at the UK office for national statistics, *Journal of Official Statistics*, 2003.

HELFENSTEIN, Xavier, 2022. The Permanent Equipment Database (BPE) - A unique and constantly moving statistical source, *Courrier des statistiques* N° N8, Insee, november 2022.

- HORVITZ, D.G., et THOMPSON, D.J., 1952. A Generalization of Sampling Without Replacement From a Finite Universe, *Journal of the American Statistical Association*, Vol. 47, No. 260 (Dec., 1952), pp. 663 – 685.
- HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative. Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. *Courrier des statistiques* N° N1, décembre 2018.
- IMBERT, Alyssa, et VIALANEIX, Nathalie, 2018. Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes. *Journal de la Société Française de Statistique*, Vol. 159 No. 2 (2018)
- JOUBERT, Marie-Pierre, 2025. Les données de téléphonie mobile - Une source de connaissance sur la population et ses déplacements. *Courrier des statistiques* N° N13, juin 2025
- JURAN, Joseph M., 1951. Quality control handbook. *Mc Graw Hill*
- KALIKA Michel, ROWE Frantz, FALLERY Bernard, REIX Robert, et RICHET Jean-Loup, 2023. Systèmes d'information et management - Le manuel de référence sur les SI, 8^e édition, *Vuibert*.
- KITCHIN, Rob, 2014. The Data Revolution – Big Data, Open Data, Data Infrastructures and Their Consequences. *SAGE Publications*. ISBN 978-1-4462-8747-7.
- KOUMARIANOS, Heidi, LEFEBVRE, Olivier, et MALHERBE, Lucas, 2024. Record Linkage: Purposes, Practices and Quality Issues, *Courrier des statistiques* N° N11, june 2024.
- LAITILA, Thomas, WALLGREN, Anders, et WALLGREN, Britt, 2011. Quality Assessment of Administrative Data. *Methodology reports from Statistics Sweden* 2011:2.
- LAMARCHE, Pierre, et LOLLIVIER, Stéfan, 2021. FIDÉLI, The integration of tax sources into social data. *Courrier des statistiques* N° N6, july 2021.
- LAWRENCE, David, et MCKENZIE, Richard, 2000. The general application of significance editing. *Journal of Official Statistics*, Vol. 16, No 3, 2000, pp. 243-253.
- LECLAIR, Marie, 2019. Using Scanner Data to Calculate the Consumer Price Index, *Courrier des statistiques* N° N3, december 2019.
- LEFEBVRE, Olivier, 2024. The Répertoire Statistique des Individus et des Logements (Statistical Register of Individuals and Dwellings, RÉSIL) – A new reference universe for demographic and social statistics. *Courrier des statistiques* N° N11, june 2024.
- LEFEBVRE, Olivier, SOULIER Manuel, et TORTOSA Thomas, 2024. The Receipt of Administrative Data: a Structural Process. *Courrier des statistiques* N° N11, june 2024.
- LEHUEDE, Sebastian, 2024. When friction becomes the norm: Antagonism, discourse and planetary data turbulence. *New Media & Society* 26 (7), 3951-3966.
- LESSLER, J.T., KASLBEEK W., 1992. Nonsampling error in surveys. New York: *Wiley*.
- LESUR, Romain, 2025. Sources de données privées : panorama et perspectives. *Courrier des statistiques* N° N13, juin 2025

- LITTLE, Roderick J.E., 1982. Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association*, Vol. 77, No. 378 (Jun., 1982), pp.237-250
- LITTLE, Roderick J.E., et RUBIN, Donald B., 2002. Statistical Analysis with Missing Data. Wiley.
- LOMBARDO, Giacomo, 2023. Data Friction in Data Sharing: a Physics Inspired Model, *Thèse en ingénierie informatique, institut polytechnique de Milan*.
- LOSHIN, David, 2011. The practitioner's guide to data quality improvement. Burlington, MA : *Morgan Kaufmann*. ISBN 978-0-12-373717-5.
- LOTHIAN, Jack, HOLMBERG, Anders, et SEYB, Allyson, 2019. An Evolutionary Schema for Using “it-is-what-it-is” Data in Official Statistics, *Journal of Official Statistics*, Vol. 35, No. 1, 2019, pp. 137–165.
- LYBERG, Lars J., WEISBERG, Herbert E., 2016. Total survey error : a paradigm for survey methodology. In *The SAGE handbook of survey methodology*, SAGE Publications.
- MALHERBE, Lucas, 2023. Appariements de données individuelles : concepts, méthodes, conseils. Insee, document de travail M2023/03, juin 2023.
- MATTHEWS, Gregory J., HAREL, Ofer, 2011. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys* Vol. 5 (2011) 1–29. ISSN: 1935-7516.
- MEIERHOFER Jürg, STADELMANN, Thilo, et CIELIBAK, Mark, 2019. Data products. In *Braschler, Stadelmann, Stockinger (Eds.): “Applied Data Science - Lessons Learned for the Data-Driven Business”*, Springer
- MEYER, Bruce D. , MOK, Wallace K. C., et SULLIVAN, James X., 2015. Household Surveys in Crisis, *Journal of Economic Perspectives*, Volume 29, Number 4, pp. 199–226
- MOREAU, Sylvain, 2024. La statistique annuelle d’entreprises : sa nature, son histoire, ses enjeux, *Courrier des statistiques* N° N12, Décembre 2024.
- NATIONAL ACADEMIES of Science - Engineering – Medicine, 2019. Reproducibility and replicability in science. *The National Academies Press* ; Illustrated edition (October 2019)
- NEYMAN, J. , 1934. « On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection », *Journal of the Royal Statistical Society*, 97, p. 558 à 625.
- NORDBOTTEN, Svein, 2010. The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries, *Official Statistics in Honour of Daniel Thorburn*, pp. 205–223
- OECD, 2008. OECD Glossary of Statistical Terms. *OECD Publishing*.
- OLSON, Jack E., 2003. Data Quality – The Accuracy Dimension. [en ligne]. Janvier 2003. Morgan Kaufmann. ISBN 1-55860-891-5.
- ÓVÁRI, Kristián, KOČIŠ, Martin, 2022. Population Census 2021 in the Slovak Republic: The “Signs of Life” method. Q2022, Vilnius.

- PERIGNON, Christophe, GADOUCHE, Kamel, HURLIN, Christophe, SILBERMAN, Roxane et DEBONNEL, Éric, 2019. Certify reproducibility with confidential data. *Science*. Juillet 2019.
- PUCHE, Véronique, 2022. Nouveau regard sur « La branche Famille et la protection des libertés individuelles ». *Regards* 2022/1 N°60, pp. 76 à 82.
- RANCOURT, Eric, 2018. *Admin-First as a Statistical Paradigm for Canadian Official Statistics: Meaning, Challenges and Opportunities. Session 3A on Wednesday, November 7, 2018, Ottawa Canada. In Proceedings of Statistics Canada Symposium 2018.*
- REDMAN, Thomas C., 1997. Data Quality for the Information Age. Janvier 1997. *Artech House Computer Science Library*. pp. 227-232. ISBN 978-0-89006-883-0.
- REDOR, Patrick, 2023. Statistical data confidentiality: a key challenge for the Official Statistical System, *ourrier des statistiques* N° N9, june 2023
- REID Giles, ZABALA Felipa, HOLMBERG Anders, 2017. Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ. *Journal of Official Statistics*, Vol. 33, No. 2, 2017, pp. 477–511
- RENNE Catherine, 2018, Understanding the Nominative Social Declaration (DSN) for Better Statistical Measurement, *Courrier des statistiques* N° N1, december 2018.
- REY, Olivier, 2016. Quand le monde s’est fait nombre. *Stock*. Octobre 2016
- RICCIATO, Fabio, 2024. Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics, *Journal of Official Statistics* 2024, Vol. 40(1) 3–15
- RIVIÈRE Pascal, 2018. Using Administrative Declarations for Statistical Purposes, *Courrier des statistiques* N° N1, december 2018.
- RIVIÈRE, Pascal, 2020. What is Data? Impact of External Data on Official Statistics. *Courrier des statistiques* N° N5, december 2020.
- ROUPPERT Benoît, 2005. Modélisation du processus de traitement d’une source administrative à des fins statistiques. *Document de travail SGI, Insee*, 2005
- SALEMINK, Irene, DUFOUR, Stéphane et VAN DER STEEN, Marcel 2020. A vision on future advanced data collection, *Statistical Journal of the IAOS* 36 (2020) 685–699
- SALGADO, David, et OANCEA, Bogdan, 2020. On new data sources for the production of official statistics. *ArXiv preprint*, 15 mars 2020.
- SCHWEINFEST, Stefan, et JANSEN, Ronald, 2021. Data Science and Official Statistics: Toward a New Data Culture. *Harvard Data Science Review* • Issue 3.4, Fall 2021
- SHAH Syed Iftikhar Hussain , PERISTERAS Vassilios et MAGNISALIS Ioannis, 2021. DaLiF: a data lifecycle framework for data-driven governments. *Journal of Big Data* (2021) 8:89, <https://doi.org/10.1186/s40537-021-00481-3>
- STATISTICS NEW ZEALAND, 2016. Guide to reporting on administrative data quality. Retrieved from www.stats.govt.nz

- SUNDGREN Bo, 1993. Statistical Metainformation Systems – Pragmatics, Semantics, Syntactics. Journal of the IAOS, Vol. 10, Issue 2. <https://doi.org/10.3233/SJU-1993-10205>
- TAVERNIER, Jean-Luc, 2020. INSEE Operations During the Lockdown Period, *Courrier des statistiques* N° N5, december 2020
- TIIT Ene-Margit, 2017. Residency testing. estimating the true population size of Estonia. *Statistics in transition new series*, June 2017, Vol. 18, No. 2, pp. 211–226
- UNECE, 2011. Using Administrative and Secondary Sources for Official Statistics. A Handbook of Principles and Practices. United Nations Economic Commission for Europe, <https://digitallibrary.un.org/record/719971?v=pdf#files>
- UNECE, 2019. GSBPM. In : site statswiki de l'UNECE. Disponible à l'adresse : <https://statswiki.unece.org/display/GSBPM>.
- UNITED NATIONS, 2025. Module for Quality Assurance when using Administrative and Other Data Sources to produce Official Statistics. Statistical Commission, Expert Group on National Quality Assurance Frameworks, mars 2025
- VAN DELDEN, Arnout, PANNEKOEK, Jeroen, BANNING, Reinder et DE BOER, Arjen, 2016. Analysing correspondence between administrative and survey data. Statistical Journal of the IAOS 32 (2016) 569–584 569. DOI 10.3233/SJI-160972
- VAN DROMME Dries, BOYDENS, Isabelle, BONTEMPS, Yves, 2007. Data quality : tools. *Rapport Smals Research 2007/TRIM3/02*. Septembre 2007.
- VICARD, Augustin, 2023. Quantifying participation in sports: a sociological and health-based approach. *Courrier des statistiques* N° N10, december 2023.
- VOLLE, Michel, 1980. Le métier de statisticien. [en ligne]. *Éditions Hachette Littérature*. ISBN 978-2-010045295.
- WALLGREN, Anders et WALLGREN, Britt, 2007. Register-based Statistics : Administrative Data for Statistical Purposes. New York : *Wiley*.
- WALLGREN, Anders et WALLGREN, Britt, 2016. Frames and Populations in a Registerbased National Statistical System. *Journal of Mathematics and Statistical Science*, Volume 2016, pp. 208-216.
- WIENER, Norbert, 1948. Cybernetics or control and communication in the human and machine, *MIT*, 1948
- ZHANG, Li-Chun, 2012. Topics of statistical theory for register-based statistics and data integration. In : *Statistica Neerlandica* (2012) Vol. 66, nr. 1, pp. 41–63
- ZHANG, Li-Chun, 2021. On provision of UK neighbourhood population statistics beyond 2021. *University of Southampton*. *ArXiv*, 4 nov 2021

Annex 1: Possible classification systems for administrative data

The term ‘administrative sources’ encompasses a variety of data which may come from very different origins, and be very different in nature. In the existing literature, we find a few attempts to establish a clear classification (UNECE 2011; G. J. Brackstone 1987). In this annex we look at three existing classification systems, and propose a fourth of our own devising.

- **System proposed by UNECE in 2011**

https://unece.org/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf

UNECE proposes a thematic approach, partly aligned with administrative structures. The handbook published in 2011 lists the following themes:

Tax data

- Personal income tax
- Value Added Tax (VAT)
- Business / profits tax
- Property taxes
- Import / export duties

Social security data

- Contributions
- Benefits
- Pensions

Health / education records

Registration systems for persons / businesses / property / vehicles

Identity cards / passports / driving licenses

Electoral registers

Register of farms

Local council registers

Building permits

Licensing systems e.g. television, sale of restricted goods

Published business accounts

Internal accounting data held by businesses

Private businesses with data holdings:

- Credit agencies
- Business analysts
- Utility companies
- Telephone directories
- Retailers with store cards etc.

- **System proposed by UNECE in 2018**

https://unstats.un.org/sdgs/files/meetings/sdg-inter-workshop-june-2018/Day2_Session3_Adm%20Data_UNSD.pdf

At a 2018 conference, UNECE presented a more aggregated approach to distinguishing between sources, with reference to the broader objectives of administrative processes:

- Data reported to administrative authorities by individual persons /legal entities for legal compliance or to access government services
- Data recording decisions made by administrative authorities
- Data generated by administrative authorities to support planning, implementation, monitoring and evaluation of administrative programmes

- **System proposed by Statistics Canada in 1987**

<https://www150.statcan.gc.ca/n1/pub/12-001-x/1987001/article/14467-eng.pdf>

Looking a little further back, Brackstone attempted a similar exercise with Statistics Canada. He looked at the stated objectives of sources, which have implications for data quality and coverage:

1. *Records maintained to regulate the flow of goods and people across borders .*

2. *Records resulting from legal requirements to register particular events*

Examples include births, deaths, marriages, divorces, business creations and mergers, licences etc.

3. *Records needed to administer benefits or obligations.*

Including tax collection, unemployment benefits, pensions, health insurance and family allocations.

4. *Records needed to administer public institutions*

Examples: documents required by schools, universities, healthcare facilities, courts and prisons.

5. *Records arising from the government regulation of industry*

For example, this includes documents relating to transport, banking, radio transmissions and telecommunications. This category also encompasses documents pertaining to price and supply management for certain products, especially in the agricultural sector.

6. *Records arising from the provision of utilities* Including electricity, telephone and water connections.

- **A new proposal**

The systems described above use different categorisation scales, with different grouping criteria: they may be concerned with identifying themes, or else with sorting data with reference to the stated goal of the administrative process or the entity which provided the data.

Some data are dependent upon real events: they may serve to record such events (births, deaths) or, on the contrary, it may be the administrative act which makes an event official (starting a business, completing a property transaction, getting married). Finally, some data correspond to administrative decisions taken by administrative entities, without any direct connection to events in real life.

There are two criteria which take on particular importance when it comes to defining the strengths and weaknesses of different types of sources, in relation to the more familiar context of surveys: the entity whence the data originated (public or not), and the strength of connection to real life events. We also work on the assumption that the way in which data are created has a substantial influence on their properties.

Hereunder we propose six types of administrative data, before looking at certain characteristics which are inextricably linked to the process whence they originated:

1. Regular declarations;
2. Occasional declarations;
3. Requests to access rights and services;
4. Management data generated by administrative activities;
5. Registers;
6. Data produced by sensors and trackers.

At this juncture, it is important to bear in mind that a dataset provided by a government body may prove to be a composite combination of these different types of data.

1) Regular declarations

These are declarations made at the request of the administrative authority, which defines and is aware of the responding population. This is a repeat exercise, with each declaration covering a clearly defined period. The data are generally gathered by means of a form created by the administrative authority. These declarations are often covered by specific regulations, and may be compulsory. The administration authorities define the concepts and value lists. The data emanating from these processes are thus clearly defined and understood by the administrative authorities, in terms of their scope, concepts, and reference dates.

This is particularly true of fiscal and social declarations.

With some declarations, the administrative authorities may pre-fill some of the information fields using data known from other sources (the “earned income” box in your tax return may be pre-filled, for example), or based on information from previous declarations (the GMBI portal simply asked respondents to update existing information in the second year of collection, 2024). The data collection processes may thus suffer from *satisficing*, leading respondents to enter approximate information, or fail to update their information.

2) Occasional, event-linked declarations

These are declarations made at the user’s initiative (the declaring party), although they may pertain to another object (birth, death of a relative, creation of a business, declaration of an incident, change of address). Generally speaking, these are events occurring in the physical world. They may predate the declaration (births, deaths), or else the act of declaration may bring something into existence (a property transaction, a marriage or partnership, a new business). The administrative authorities create the necessary forms, and thus retain control over the structure of the data collected.

The reporting periods of such declarations is beyond the control of the administrative authorities, as the events in question are not predictable and they may only be declared some time after the event. The population concerned is not known *a priori*, and non-declaration may be an issue. We can look in greater detail at the population “authorised to make declarations” (for example, we know that births must be declared by individuals), but the declaring party and the object of the declaration may not be one and the same. We may also note that certain events are (in theory, at least) unique, involving the creation or existence of a new object, or else the termination or disappearance of an object. Other events correspond to a change of status, and may occur multiple times for the same object.

3) User requests submitted to the administration

These often correspond to applications for rights, benefits, services etc. They include requests for pensions and social benefits (allocations, subsidies and grants). The population is defined as the number of applicants. Those who are entitled to such benefits but do not request them remain outside the scope of the administrative process, although they could still be units of interest for statistical purposes. The issue here is not non-declaration, as seen above, but rather non-uptake.

For these first three categories, administrative data stem from interactions between users and administrative entities, and the process of gathering information resembles a process of observation.

4) Management data generated by administrative activities

Some forms of data are generated by the internal operations of administrative agencies. They may, for example, constitute a record of the activities of the judicial administration (Chambaz, 2018). The objects of interest to the administration may not necessarily hold the same level of interest for statisticians. Administrative objects can be difficult to define and track over time. In theory, administrative objects are clearly defined, but they may be complex (cf. the section devoted to these objects in Chapter 2).

When dealing with management data, the administrative clerk must decide when to update information: the evolution of these data is not always driven by interactions with users (career updates for pension, opening and closing of job opportunities). These data do not necessarily correspond to observations; in this case, they stem from administrative decisions.

We may find ourselves with a mixture of observation data and management data for the same administrative object. Depending on the structure of the administration, management practices may vary from one entity to the next.

We might also include in this category some very specific types of data: the functional parameters of administrative entities (pricing, pay scales etc.), which are in reality “high level” data used to calculate many other data points.

5) Registers

Registers occupy a slightly different position compared with the types of sources listed above. In some respects, they do not really constitute a separate category. Their purpose is to make reference data available for a variety of users. The data in question may have originated from declarations, coordinated by the administration in order to corral information from different time frames into a single dataset.

Registers, to the extent of they are *lists of instances of a given entity* (Rivière, 2022), seek to provide the most exhaustive possible description of objects of interest to the administration, recording (as far as is possible) any changes of status or attributes pertaining to these objects.

Just as datasets may contain a mixture of declaration data and management data, some datasets may combine information derived from administrative registers with recent declaration data (examples: tax data, CAP data).

6) Data produced by automatic measurement systems

This final category contains the least frequently utilised types of sources. They include physical sensors (the GPS chips installed on ships, or use to measure road traffic, or meteorological conditions) as well as digital trackers (bank card data, use of health insurance cards).

What emerges clearly from this attempt at categorisation is that the notions of “object”, “data producing entity” and “time frame” are important in the construction, and subsequent use, of data.

Whatever type of administrative data we use, statistical operations are likely to run into certain difficulties: administrative concepts and classifications diverge from the units of interest to statisticians, mechanisms for handling non-responses or coverage problems with the theoretical scope are not recorded, and, on a more general level, our knowledge of the data collection and processing operations is often limited.

Furthermore, data depend upon the existence of the processes which they support, and these processes are liable to change drastically, or even disappear, as witnessed by the example of council tax in France.

Nonetheless, the origin of the data also explains certain differences: the existence of a formally defined data gathering protocol means that administrative data may more closely resemble survey data.

Table: Attributes of the different types of administrative data

	Initiating party	Collection process			Prior knowledge of the reference population
		Clearly defined data collection format	Clearly defined collection period	Reference period	
Regular declaration	Administration	Yes	Yes	Yes	Yes
Occasional declaration	User	Yes	No	Yes	No
User request	User	Yes	No	Yes	No
Internal purposes	Administration	No	No	No	No
Automated measurement system	Administration	Yes	No	Yes	Yes

Annex 2: Dimensions used by Istat to evaluate the quality of administrative data

Source	Metadata	Données		
		Dataset	Objects	Variables
		Technical checks	Accessibility of the file	Compliance of data to definitions (metadata)
		Integrability	Comparability of AD objects and statistical units (or capacity to transform objects)	Presence of comparable linking variables
		Accuracy	Units presenting inconsistencies or anomalies	Units presenting inconsistencies or anomalies
		Completeness	Overcoverage, undercoverage, duplicates	Missing values, imputed values
		Time-related Dimension	Time lag after reference period	Evolution of units over time (entering and exiting the scope) Stability of definitions

(Cerroni et al., 2014)

Annex 3: List of administrative data sources cited in this article

The list of administrative sources hereunder contains the main sources referenced in the document, bearing mind the difficulties inherent to the very notion of what constituted a source, as discussed in Chapter 3c. They are presented here by theme.

Fiscal data:

TH: *taxe d'habitation* / council tax. Until 2022, all homes were liable for council tax, to be paid by the occupants. Since 2023, council tax is only payable on second homes and holiday homes.

GMBI: *Gérer mes biens immobiliers* – Manage my properties, a fiscal data source for residential properties. Owners must declare their properties, and any occupants.

POTE: tax declaration for personal income.

Business tax declarations in France must include the annual profit and loss accounts, but also the balance sheet including fixed assets and amortisation.

Land register (or MAJIC – updated land register information). A register of all plots of land in the country, with or without buildings. Also contains information on the nature of buildings, premises and their owners.

Social data:

The DSN (*Déclaration Sociale Nominative* - Individual Social Declaration) is a “monthly file produced from pay data, intended to provide the relevant organisations and administrative bodies with the information required to manage the social protection of employees.”

PASRAU:

The PASRAU (*Prélèvement à la source pour les Revenus Autres* - Pay as you earn – Other Income) system is part of the government’s efforts to simplify and streamline social declarations. It is a logical extension of the DSN, which in recent years has significantly simplified the declaration procedures for wages and other sums paid by employers. This system, based on the NEORAU standard, is a complement to the DSN (based on the NEODES standard) for “income replacement” revenue.

DRM (*Dispositif de Revenus Mensuels* - Monthly Income System):

The DRM comprises three data processing operations: a database updated monthly with data from the individual social declarations (DSN), a database updated monthly with data for “pay as you earn sums collected but not covered by the individual social declaration, or constituting income replacement” (PASRAU) and a reporting service. This system centralises virtually all data relating to the income received by the population known to the fiscal authorities.”

CNAF: *Caisse nationale des allocations familiales* - National Family Allocations Fund

CNAV: *Caisse nationale d'assurance vieillesse* - National Old Age Insurance Fund

RGCU: *register général des carrières unique* – Career Register

Other data sources mentioned:

Vehicle registration database: a register of all road vehicles and their owners.

CAP: Common Agricultural Policy: data regarding applications for farm subsidies. Provided by the services and payments agency (ASP).

Annex 4: List of acronyms – only in english version

CAP : Common Agricultural Policy - see also annex 3

CNAF : Caisse nationale des allocations familiales, or National Family Allocations Fund - see also annex 3

CNAM : Caisse nationale d'assurance vieillesse, or National Old Age Insurance Fund - see also annex 3

CNAV : Caisse nationale d'assurance maladie, or National Health Insurance Fund - see also annex 3

CNIL : National Commission for Information Technologies and Civil Liberties (CNIL)

CNIS : National Council for Statistical Information (CNIS, Conseil national de l'information statistique)

CSG : The contribution sociale généralisée (CSG) and the contribution au remboursement de la dette sociale (CRDS) are taxes designed to finance social protection. The remuneration base subject to tax may differ from the employee's actual remuneration. DSN contains different remuneration bases

CSNS : The Non-Significant Statistical Code (CSNS) was defined by the Law for a Digital Republic of 2016 in order to allow the implementation of file matching for statistical purposes without using the NIR (or national insurance number), thus ensuring a high level of personal data protection

DGFIP : public finance directorate

DPAE : Déclaration Préalable à l'Embauche - Declaration prior to hiring

DREES : DREES is the french national statistical office of Health minister

DRM : (Dispositif de Revenus Mensuels - Monthly Income System) - see also annex 3

DSN : (Déclaration Sociale Nominative - Individual Social Declaration) is a “monthly file produced from pay data, intended to provide the relevant organisations and administrative bodies with the information required to manage the social protection of employees.” - see also annex 3

Fideli : Fichier Démographique sur les Logements et les Individus (Demographic File on Dwellings and Individuals, FIDÉLI) ile generated through the management of housing tax

Fidelimmo : project that aims to build a new statistical database, using administrative data sources and systematically combining information on households and on the real estate assets they own

Filosofi : French acronym based on Fichiers Localisés SOciaux et FIScaux, meaning localised tax and benefits files

FIP : Fichier d'imposition des personnes (Tax File for Individuals)

France travail : France Travail (English: France Employment) is a French governmental agency which registers unemployed people, helps them find jobs and provides them with financial aid

GIP-MDS : French Public Interest Group (GIP) ‘Modernisation of Social Declarations’. Created in 2000 to meet the need to simplify administrative procedures for companies, GIP brings together all the players in the social protection sector

GMBI : Gérer mes biens immobiliers – Manage my properties, a fiscal data source for residential properties. Owners must declare their properties, and any occupants. - see also annex 3

MSA : MSA (Mutualité Sociale Agricole) provides social security coverage for all farmers and their dependents in France

NAF : French classification of economic activities

NIR : NIR is the national insurance number, and the identification number in the National Register for the Identification of Individuals

PASRAU : (Pay as you earn – Other Income) data source, corresponding to declarations of “replacement” income sources such as unemployment benefits, pensions etc - see also annex 3

POTE : tax declaration for personal income - see also annex 3

Résil : Répertoire statistique des individus et des logements (Statistical Register of Individuals and Dwellings)

RGCU : register général des carrières unique – Career Register - see also annex 3

RNCPS : Répertoire National Commun de la Protection Sociale (Common National Social Security Register) is a register used to combat fraud

RNIPP : Répertoire national d’immatriculation des personnes physiques (National Register for the Identification of Individuals, RNIPP)

RSA : The Active Solidarity Income (RSA for Revenu de Solidarité Active) guarantees a minimum level of income for people without resources

SIREN : SIREN is the identification number in the business register SIRENE

Sirene : Système Informatique pour le Répertoire des ENTREPRISES et des Établissements (Computer system for the business and establishment register)

Sirus : Système d’immatriculation au répertoire des unités statistiques (Statistical Business Register) is linked to SIRENE and meets the requirements of the statistician: it takes into account the notion of enterprise in the economic sense, it integrates new units such as groups of companies, and it implements specific processing such as statistical cessation

TH : taxe d’habitation / council tax. - see also annex 3

URSSAF : The "Union de recouvrement des cotisations de sécurité sociale et d’allocations familiales" (Urssaf) is a French organization. Its main missions are to collect employee and employer social security contributions to finance social protection (pension funds, unemployment insurance, etc.)