The Receipt of Administrative Data: a Structural Process

Olivier Lefebvre *, *Manuel Soulier*** and *Thomas Tortosa****

Given the growing importance of administrative data in the statistical production process, rationalising the way it is processed is becoming a major challenge.
The work starts as soon as the data arrive! They are received as configured according to their administrative uses, thus requiring to transform them into statistical data, i.e. data organised according to the statistical units of interest (individuals, households, establishments, employers, etc.) and statistical concepts. This phase can be pooled and "decoupled" from downstream statistical processes, which provides greater scope for adaptation, but also for sharing information. The challenge is to set up a system that combines adaptability, performance, security and traceability.
INSEE has adopted this approach with the ARC (Accueil-Réception-Contrôle – receipt, acceptance, control) tool, based on a demanding use case: the monthly processing of around 2.5 million nominative social declarations. By gradually extending its functions and performance to adapt to new data and new constraints, ARC is now an essential part of INSEE's statistical production system.

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

** IT project manager for the ARC (Accueil Réception Contrôle - Receipt, Acceptance, Control) application, Centre-Val de Loire Regional Directorate, INSEE.
manuel.soulier@insee.fr

*** Statistical project manager for the “receipt of sources” project within the RÉSIL programme, DSDS, INSEE.
thomas.tortosa@insee.fr

Industrialising the integration of administrative data into our information systems is essential, given the growing importance of this type of data in our statistical production processes (Cotton and Haag, 2023). To meet this challenge, the solution adopted by INSEE is a structure for the receipt of sources that is made available to producers of statistics, based on a modern and shared generic tool.

When administrative data was initially being used, this integration work was completed separately for each source: each party developed its own process, adapted to the source being processed and to downstream processing operations. This model worked for decades. However, during the 2010s, this administrative data became more numerous, more frequent, more scalable, and more capable of serving as input in multiple data production lines. In addition, new needs have emerged in the reception process which needed to be more "adaptable", efficient, traceable and open, while remaining highly secure.

The Expected Qualities of an Administrative File Receipt Service in a Statistical Universe: Adaptability, Performance, Traceability and Security

The service must be **adaptable**. It must take changes in the content or format of the information provided into account as soon as possible. Such transformations are inevitable, because, as with public policies or their implementation processes, the administrative data are not static and change to reflect the needs of the public policy they accompany.

It must be possible to apply these changes quickly, without the changes being propagated simultaneously to all files if they come from a range of different bodies. Multiple versions of the files to be received can coexist, with content modified on the basis of the creation date.

To meet these needs for adaptability and responsiveness, the system must simultaneously manage and receive several versions of files while offering "generic" processing, that is, processing that works on all files.

> ***Statisticians must have tools allowing them to configure the processing parameters and measure the impact of parameter changes on the data.***

As part of the receipt function, statisticians must also have tools allowing them to configure the processing parameters and measure the impact of parameter changes on the data.

In addition, this function must not only receive files with formats and content that can change rapidly, but it must also feed into statistical production chains using data that is stable over the long term. Extending the principle of ensuring the generic nature of the tool used as an interface between receipt and client applications is one solution to address this problem.

Computer **performance** becomes crucial. This is because data are transmitted more and more frequently and must be assessed quickly. Extremely cumbersome sources are now received every month by INSEE and need to be processed quickly to ensure the relevance of the statistics produced within increasingly short deadlines (European social demand).

Traceability is a fundamental quality requirement in a production process. The more changes there are, the more essential it is to be able to trace them. This makes it possible, if necessary, to reproduce the processing, to report on the operations carried out and thus to more easily analyse any changes to be made to the downstream processing and, ultimately, to document the process.

Administrative data are an important data source. They have long been feeding into statistical production processes but could be used to test new statistical processing operations. With the new data science professions and the increasing number of data scientist roles within INSEE, it is necessary to provide mechanisms that enable controlled access to raw statistical data[1] for the purpose of using them innovatively.

Lastly, **security** is essential for this type of data, which may contain personal data that require strict confidentiality, while their integrity must also be protected; the traceability requirement mentioned above also contributes to the overall security of the process and processed data.

Establishing the cross-functional receipt of data therefore requires the centralisation of security rules within the framework of administrative data governance. For example, this involves pooling and carrying out cross-functional processes, such as the "pseudonymisation" of data, as soon as possible (Cotton and Haag, 2023) and implementing a policy and tools for managing access rights, while also leaving open the possibility for each owner to apply additional specific rules.

> ***This had led to the pooling of an administrative data receipt tool.***

These requirements necessitate significant investments; this had led to the pooling of an administrative data receipt tool that is capable of managing sources of different natures and origins and feeding into various operating processes. Such a tool is based on the decoupling of the data receipt function and the data processing or analysis function.

[1] Raw statistical data are administrative data that exist in a statistically usable format. Statisticians consider these data to be "raw" because they have not yet been processed for statistical use.

Decoupling the Data Receipt Phase from the Statistical Processing Phase...

Creating the data receipt service means looking at the receipt phase as an inherently separate activity, which must be distinct from the processing operations necessary for the creation of a statistical product.

Traditionally, the process of developing a statistical product based on external data is iterative in nature. Following an initial appropriation stage, the statistician integrates their file to obtain an expected result by performing various stages. While the stages in question may vary, they may include the following:

- if necessary, structuring the file and transforming it into a database;
- renaming the variables in order to ensure the sustainability of the processing or explaining the names of the variables;
- reprocessing the variables in order to correct certain imperfections (non-responses, outlier values, etc.);
- creating statistical variables derived from one or more variables that are sometimes modified or aggregated (all those that make up wages, all the income of a category of workers, etc.);
- making a product for dissemination.

These phases are mostly carried out in blocks.

... Pooling the Receipt of Data...

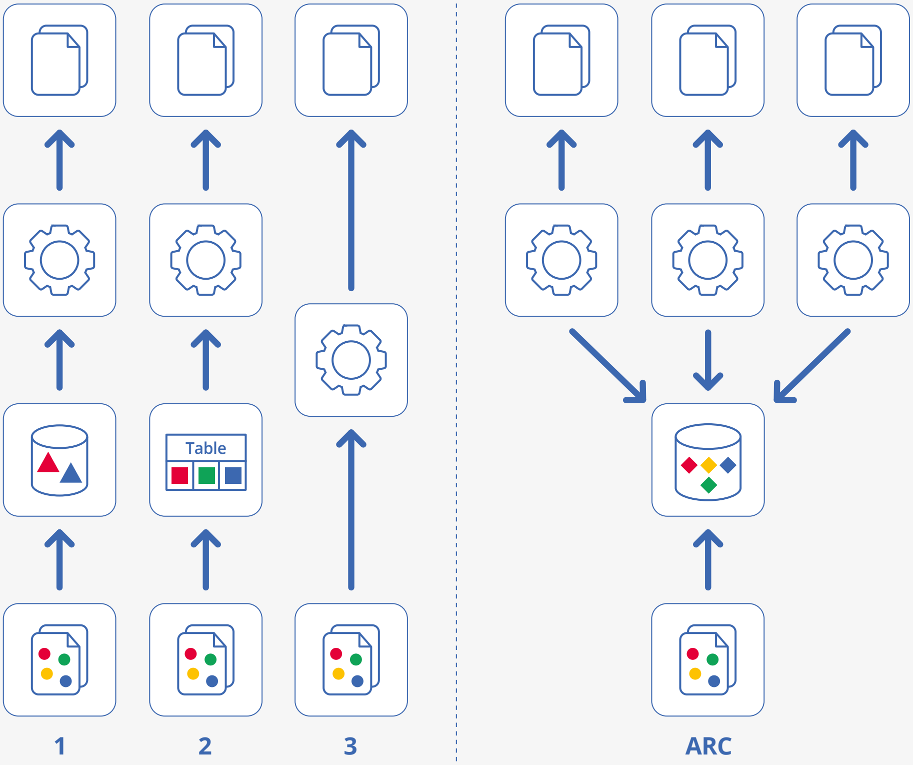
Statisticians have all the data contained in the file and, in stages, they will transform them into a statistical product. However, while the process is easy during the creation stage, it quickly becomes difficult to maintain and use if it is not developed and implemented in a modular way. When transforming data, modifying the production chain can quickly become complex if the processing phase is closely linked to the reception phase, i.e. if the processing operations are based directly on the raw administrative data.

The idea of decoupling the receipt of data has gradually emerged (***Figure 1***). For example, INSEE set up the receipt of the Déclaration Sociale Nominative (Nominative Social Declaration, DSN)[2], and developed a dedicated tool (Accueil-Réception-Contrôle [Receipt, Acceptance, Control], known as ARC), to feed into the production of statistics on employment and earned income (Renne, 2018). This is also the case for the Élaboration de Statistiques Annuelles d'Entreprises (Elaboration of annual statistics of companies, ESANE) process, which is based on both survey data

[2] The Déclaration Sociale Nominative (Nominative Social Declaration, DSN) is a mandatory, unified and online declaration made by each employer. The DSN makes it possible to ensure the recovery of social security contributions and to allocate entitlements to employees.

and tax data. More recently, the project to redesign FIDÉLI[3] following the abolition of housing tax, split the application into two parts: the first is for the receipt of the data and the second is to perform the processing operations. Carried out after the experiments, the decision to separate the receipt of external sources from statistical processing operations is an extremely useful one for statisticians.

▶ Figure 1 - Receipt, the First Step in Rationalising Processing Operations



Reading note: The data are received repeatedly. Each pipeline manages the data differently as they arrive. Pipeline 1 selects the useful data and modifies not only their form but also their content, which it stores in a database. Pipeline 2 selects the data and refines them, before storing them in a table that can be used by statistical software (e.g. SAS or R). Pipeline 3 integrates the file as a whole into these processing operations.

Reading note: The ARC receipt service restructures the information in the file, but does not change it. Each pipeline can examine the file to retrieve the data.

[3] FIDELI: Fichier Démographique d'origine fiscale sur les Logements et les Individus (Housing and individual demographic files) (Lamarche and Lollivier, 2021).

… to Ensure Rationalised and Controlled Use!

With this decoupling, the information system is more robust. The receipt of sources largely absorbs exogenous shocks, if the service was designed with this in mind[4]. The removal, addition or modification of information can be processed in this phase in order to feed into the downstream statistical processing operations in an almost identical way, minimising maintenance of the latter. ARC is based on this principle: a data design phase, that is, a phase in which the received data are transformed into a system of usable raw statistical data. This phase makes it possible to manage changes, such as renaming variables or modifying their content, when possible.

For example, the POTE[5] file of income tax returns from the DGFiP[6] is in text format; therefore, reading is only possible thanks to a file layout.

Position		Lg	Numeric or alpha	Reading format	Writing format	Input format	PACname	Description
247	254	8	9(8)	8.		8.	DADOKZ	SITFAM: DATE OF DEATH OF 2042

Here, the variable DADOKZ, which corresponds to the date of death of the fiscal reference person of the household, can be read in the file from position 247 to 254 (length 8). The format specified is a numeric format (length 9), while the nature of the variable is of the date type.

At the end of the receipt process, the data will be made available in a date_dc variable (which explains the content of the variable) in date format ("YYYY-MM-DD").

Decoupling makes multiple uses of data possible and simpler. When data receipt is integrated into a production chain of a statistical production application, it is difficult, if not impossible, to open up this access to data to other applications.

For example, the DSN data loaded into ARC were initially intended for the structural chain for calculating salaried employment.

This produces annual employment statistics in terms of numbers or distribution, by status or economic activity. For this use, the structural chain is the "client" and the consumer of the DSN data received in ARC.

When INSEE wanted to use the same data to make cyclical estimates of salaried employment, it was easier to adapt due to decoupling: like the structural chain, the new cyclical chain was declared a "client" of ARC. This would have been almost impossible if the receipt of the DSN had been integrated and coupled with the structural chain.

[4] See the section hereinafter on new professions.

[5] Fichier Permanent des Occurrences de Traitement des Émissions (Permanent File for Occurrences of Processing of Issues): it contains the data relating to the tax returns for the year sent by taxpayers to the DGFiP in the spring of the following year.

[6] DGFiP: Direction générale des Finances publiques (Public Finances Directorate General) is a directorate within the French central general government, which reports to the Ministry of the Economy, Finance and Industrial and Digital Sovereignty.

> ***Isolating the receipt process provides an opportunity to manage data access rights for client applications.***

In addition, isolating the receipt process provides an opportunity to manage data access rights for client applications, each of which can select the data it needs from the data made available. Sharing is implemented at source level without any need to build a "gateway" between applications.

Lastly, this decoupling also makes it possible to identify the receipt of data as a separate process, with all the advantages that entails. It is thus possible to decouple it from client processes and therefore to change it independently, or to add an additional client process fed into at source level. This also makes it possible to "target" examinations and investments to optimise it.

▶ Managing and "Activating" Metadata

> ***The data made available cannot be used without metadata.***

The data made available cannot be used without metadata: they are taken from a phase for the transformation of administrative data into statistical concepts. The metadata are used to document the data produced and are referred to in such cases as "passive". They are generated during processing and make it possible to keep a record of the operations carried out on the data, as well as the use of variables, both internally and externally. To that end, INSEE has a statistical metadata repository, RMéS, which enables metadata to be managed, shared and disseminated (Bonnans, 2019).

In the field of surveys, metadata are used as an input for the process of designing the collection medium in order to generate it (Cotton and Dubois, 2019). This use is possible thanks to the establishment of a set of tools and services related to RMéS. In such cases, they are referred to as "active" metadata, meaning that they have a function within the statistical process beyond a documentary function. One of the challenges in the receipt of administrative data is therefore to make these metadata "activatable", such as is done with surveys; this involves documenting the metadata from administrative files as early in the process as possible.

Providing producers with the metadata associated with the raw statistical data gives them what they need to document their process and ultimately integrate their own metadata into RMéS. A recent experiment on land data has shown that most of the metadata entered in the receipt phase could be reused without modifications in subsequent processes, up to the creation of the databases for dissemination.

In addition, delivering metadata as an input for the processing process makes it possible to establish “production metadata”, which enable subsequent modifications to the data to be traced, configured or even specified, unlike “dissemination metadata”. This is referred to as data lineage[7] (Biseul, 2023).

▶ Deux nouveaux métiers : modélisateur de données et intendant des données

Establishing a receipt service requires special adaptation of this data receipt phase. When we talk of special adaptation, we must talk about specialist professions. Two new professions have emerged in connection with the various tasks to be carried out.

The first of these is a **data modeller**. Data modellers design the models into which the administrative data for users will be inserted. They transform a given model, over which they have no influence, into a statistically usable model; the latter must be both robust to changes and constructed in such a way that users can easily use the data to produce statistics. Not only must they therefore be skilled in modelling, but they must also be attentive to users to ensure that the model developed aligns with expectations.

As the data are administrative in nature, it is sometimes difficult to transform them into a statistical concept. Modellers can segment them by theme: this is then referred to as vertical partitioning of the data. Technical modelling then completes the semantic modelling.

For example, the file of income tax returns (POTE), provided by the DGFiP, is a very large file, both in terms of the number of lines (45 million) and the number of variables (600 for the fixed part, more than double that for the variable part). Within that file, multiple themes related to the general topic of taxation can be identified: income tax, the general social contribution, tax on property assets, etc., all of which can be isolated. By nature, the existence and characteristics of a tax are not permanent: housing tax is an example of this. The advantage of this modelling is that it makes it possible to build a model around themes for each tax. Ultimately, the use of the file will be more robust since only the parts affected by the changes are modified.

The second new profession is a **data steward**. Data stewardship is a concept with multiple definitions, each with a remit that varies. The definition used in this article is the one used by Statistics Canada[8] specifically: “Data stewardship is data governance in action, i.e. the operational implementation of data policy.” It is therefore the effective implementation of the rules governing the collection, management, security, quality and dissemination of data within an organisation.

Thus, for data stewards, data administration skills are required since they ensure the receipt, control and documentation of the data they make available.

[7] *https://www.journaldunet.fr/web-tech/guide-du-big-data/1516833-data-lineage-definition-principes-et-outils/.*
[8] *https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062020013.*

They must also have data management skills, as they manage data access rights and monitor agreements concluded with suppliers (respect of deadlines and data transmission and retention formats).

Regarding access to the data, the plethora of potential uses for the same administrative data means that their security becomes a more complex issue. Once “pseudonymised” and transformed into a statistical format, the data are more open and more “shareable”. However, this sharing must be both organised and selective, based on specific purposes and in accordance with the principle of proportionality of processing. Data stewards must apply the rights of users in relation to their access to the data.

Lastly, these specialists are in direct contact with users and producers of administrative data. In particular, they are the first port of call for resolving data transmission problems. Therefore, interpersonal skills are also required.

Data stewards are responsible for the collection and management of the data, but not for their uses. It is therefore necessary to maintain a relationship between the producer of the statistics and the supplier, which is unburdened by management issues and is therefore focused on the data’s content, use and also evolution.

▶ Accueil – Réception – Contrôle (Receipt, Acceptance, Control - ARC), INSEE’s Computerised Data Receipt Service –

> ***The ARC IT application has been providing the receipt service at INSEE for around a decade.***

The ARC IT application has been providing the receipt service at INSEE for around a decade, since it first existed in 2015 with the receipt of the DSN files.

ARC functionally covers some of the phases of the GSBPM[9] (*Figure 2*). This is a standard framework for statistical organisations that allows them to adopt common terminology to describe the life cycle of a statistical operation (Erikson, 2020).

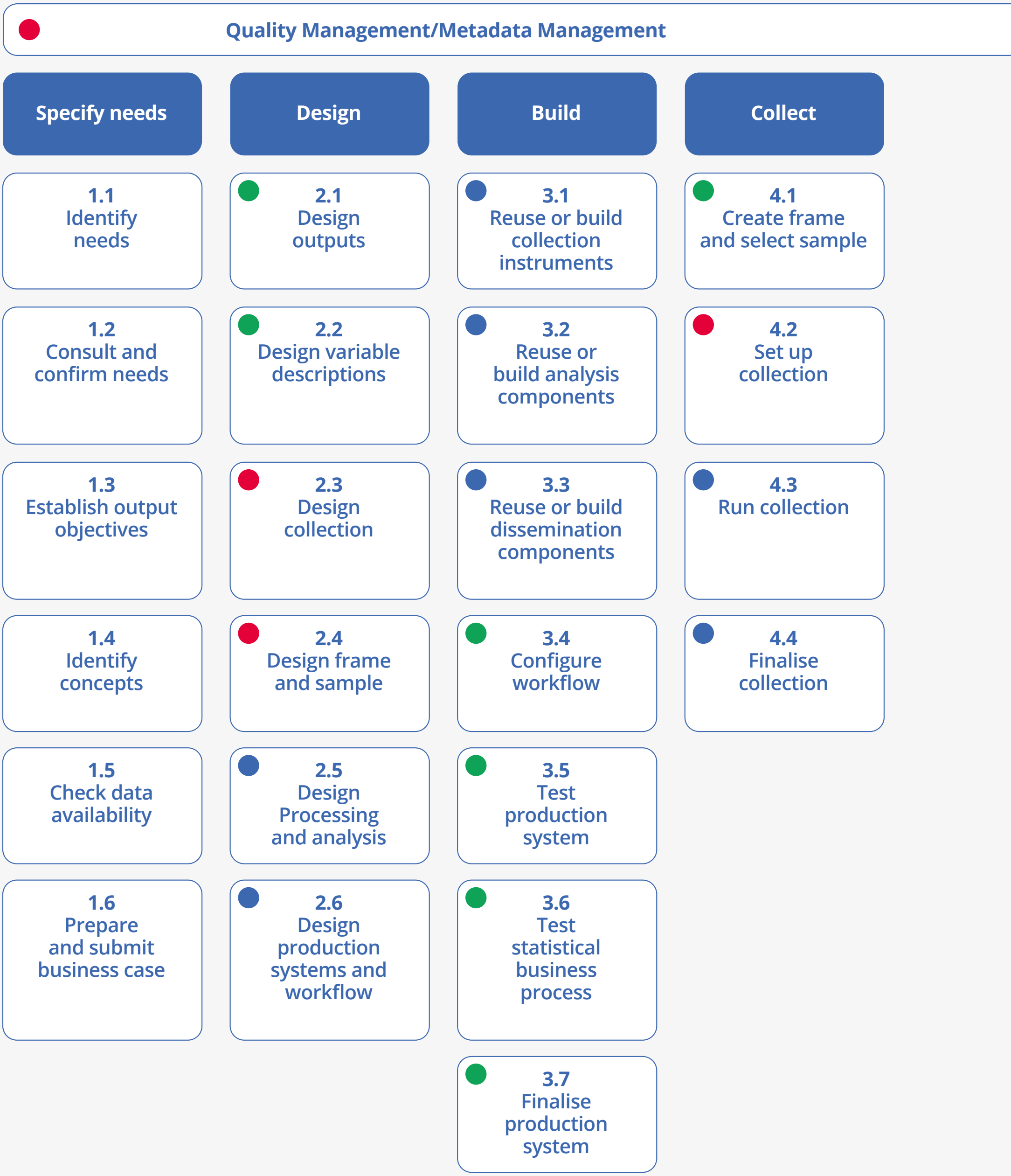
Statisticians rely on the cross-functional metadata management operation to create the design of the final dissemination products and the design of the description of the variables used. To that end, they directly use the implementation offered in the ARC application or submit DDI[10] modelling (Dondon and Lamarche, 2023) to it, created using the Colectica Designer[11] tool.

[9] The Generic Statistical Business Process Model (GSBPM) describes the various stages to follow when producing official statistics.

[10] DDI: the Data Documentation Initiative is an international consortium of research institutes and producers of statistics that aims to define standards for the documentation of statistical data, with a particular focus on survey data, collection methods and repositories (nomenclatures, codifications, etc.) used for collection. The DDI format is based on the XML format (see the definition of the XML format hereinafter).

[11] *https://www.colectica.com/software/designer/.*

Figure 2 - The Functional Coverage of ARC Compared to the Generic Statistical Business Process Model (GSBPM)

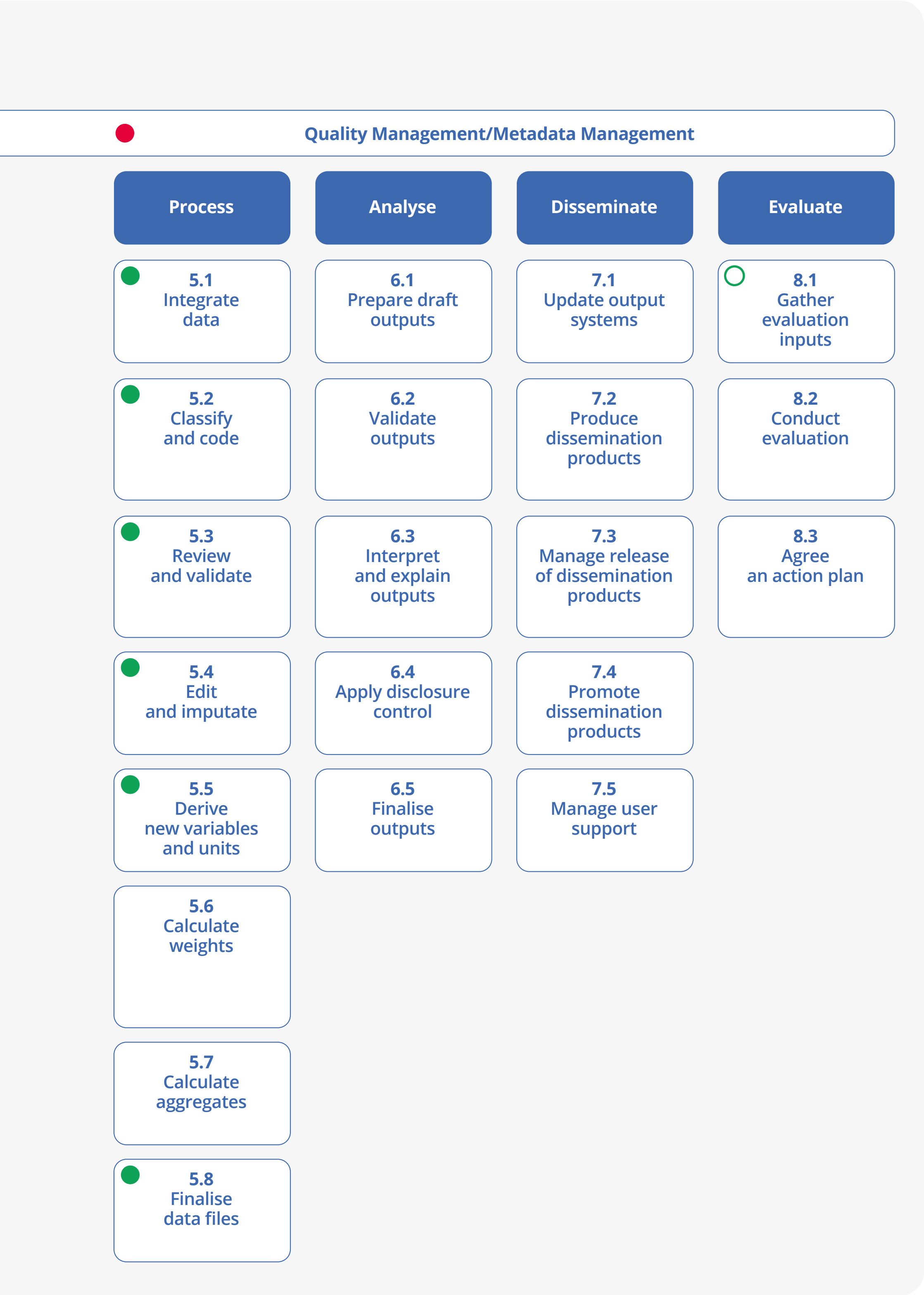


🔴 Quality Management/Metadata Management

Specify needs	Design	Build	Collect
1.1 Identify needs	🟢 2.1 Design outputs	🔵 3.1 Reuse or build collection instruments	🟢 4.1 Create frame and select sample
1.2 Consult and confirm needs	🟢 2.2 Design variable descriptions	🔵 3.2 Reuse or build analysis components	🔴 4.2 Set up collection
1.3 Establish output objectives	🔴 2.3 Design collection	🔵 3.3 Reuse or build dissemination components	🔵 4.3 Run collection
1.4 Identify concepts	🔴 2.4 Design frame and sample	🟢 3.4 Configure workflow	🔵 4.4 Finalise collection
1.5 Check data availability	🔵 2.5 Design Processing and analysis	🟢 3.5 Test production system	
1.6 Prepare and submit business case	🔵 2.6 Design production systems and workflow	🟢 3.6 Test statistical business process	
		🟢 3.7 Finalise production system	

Key:

- 🔴 These phases are “prerequisites” for the proper functioning of ARC, to be managed both upstream and externally.
- 🔵 These phases are covered by ARC and statisticians are unable to configure them.
- 🟢 These phases are the ones that statisticians can configure in ARC.
- ○ Phase under development.



● Quality Management/Metadata Management

Process	Analyse	Disseminate	Evaluate
● 5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	○ 8.1 Gather evaluation inputs
● 5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
● 5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
● 5.4 Edit and imputate	6.4 Apply disclosure control	7.4 Promote dissemination products	
● 5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
5.6 Calculate weights			
5.7 Calculate aggregates			
● 5.8 Finalise data files			

The phases for the design of the data collection, the frame and the sample are outside the scope of the ARC application and are performed in agreement with the data provider.

The functions covering the processing and analysis design and production system and workflow design phases, as well as the three development phases, model the receipt function in ARC and provide the framework for possible configurations for statisticians. They were created during the design of the application and form the ARC pipeline (***Figure 3***).

However, statisticians are able to configure the workflow for the phases of the "processing" stage; they can test their configurations on small volumes of data, in dedicated spaces which are separate from production spaces and called "sandboxes". The sandboxes in ARC allow the production system to be tested using a reduced number of source files. When the update is complete, statisticians apply their configurations to the actual file flow and finalise the production system.

In the context of statistical processing using a receipt function, ARC covers the "design", "build" and "collect" stages of the GSBPM and shares certain phases of the "process" stage with the applications using the data. For example, ARC plays a role in the data editing and imputation phase for the publishing of certain data in order to make them statistically usable (correction of categories, compliance with the format or expected values, etc.), whereas statistical transformations in the context of specialist professions (imputations of missing values or detection of outlier values) are performed by the data processing applications.

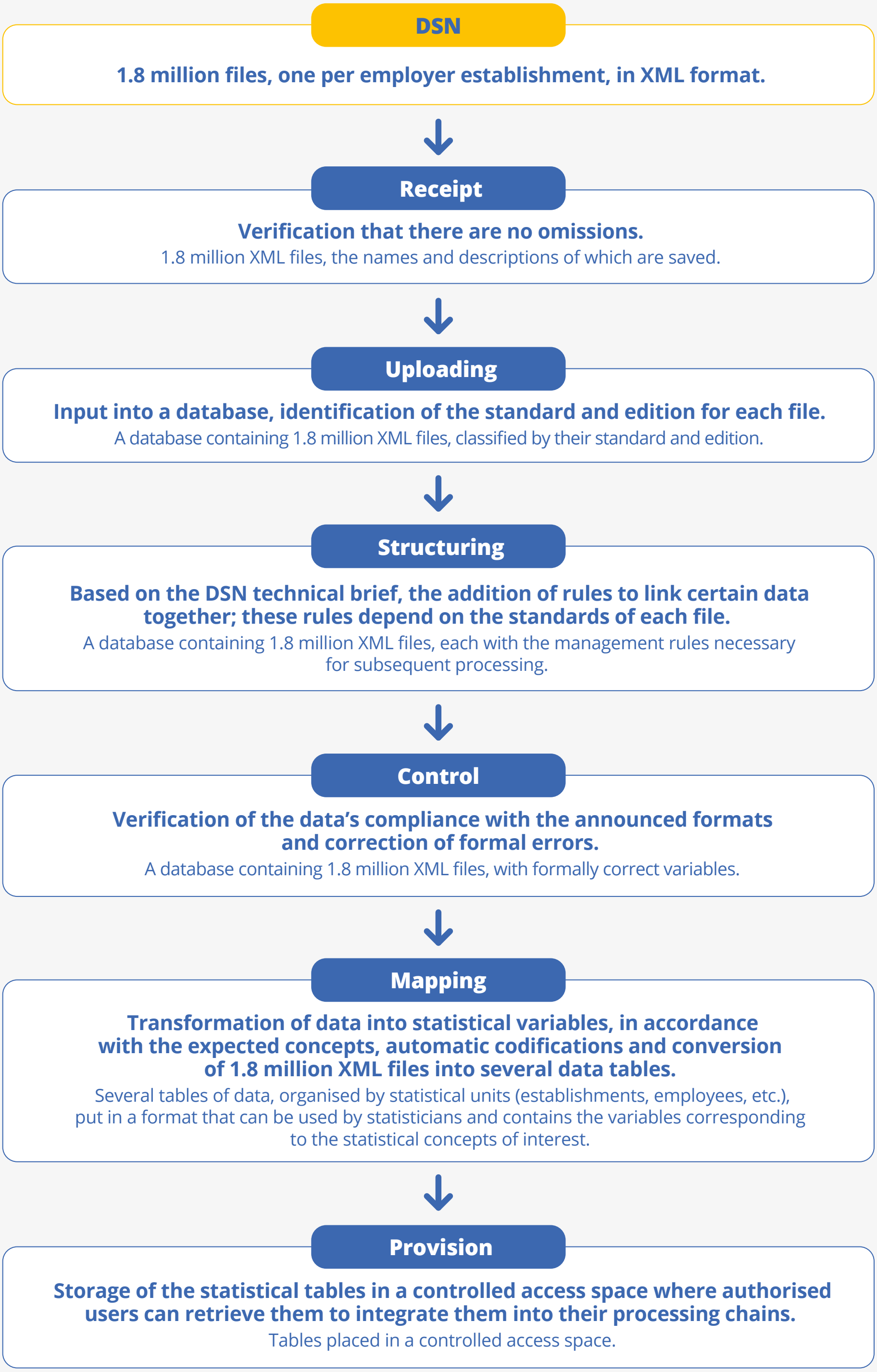
At present, there is no automated production of a quality assessment of the receipt of a source (number of records read, number of erroneous values, etc.), which corresponds to the evaluation product collection phase of the GSBPM. This is currently handled by the client's information system (IS). The ARC product will need to develop in order to offer implementation of this collection for the processing operations that concern it (and, in particular, compliance with the announced standards), as quality assurance is an essential component of the receipt function.

The Processing Pipeline Using the DSN as an Example

> ***The Nominative Social Declaration, DSN are converted into statistical data by ARC computer processing, through a succession of functional modules. Each module performs a specific operation.***

The DSN is a mandatory online declaration, allowing employers to send information about employees to social security bodies (Humbert-Bottin, 2018). In December 2023, INSEE received approximately 2.5 million files from employers containing their employees' salary data. It received 1.8 million such files in 2016.

▶ Figure 3 - The Déclaration Sociale Nominative (Nominative Social Declaration, DSN) Pipeline



DSN

1.8 million files, one per employer establishment, in XML format.

Receipt

Verification that there are no omissions.
1.8 million XML files, the names and descriptions of which are saved.

Uploading

Input into a database, identification of the standard and edition for each file.
A database containing 1.8 million XML files, classified by their standard and edition.

Structuring

Based on the DSN technical brief, the addition of rules to link certain data together; these rules depend on the standards of each file.
A database containing 1.8 million XML files, each with the management rules necessary for subsequent processing.

Control

Verification of the data's compliance with the announced formats and correction of formal errors.
A database containing 1.8 million XML files, with formally correct variables.

Mapping

Transformation of data into statistical variables, in accordance with the expected concepts, automatic codifications and conversion of 1.8 million XML files into several data tables.
Several tables of data, organised by statistical units (establishments, employees, etc.), put in a format that can be used by statisticians and contains the variables corresponding to the statistical concepts of interest.

Provision

Storage of the statistical tables in a controlled access space where authorised users can retrieve them to integrate them into their processing chains.
Tables placed in a controlled access space.

The DSNs are converted into statistical data by ARC computer processing. This processing is structured as a "pipeline", i.e. it consists of a succession of functional modules, with each module performing a specific operation. The ARC modules are pre-configured by statisticians. ARC follows a document-style approach: each file is processed individually and independently of the others and its name is retained as an identifier for each piece of data.

Integrating Administrative Data

The GSBPM data integration phase is covered in ARC by the first two modules of the processing pipeline.

In the **receipt** module, the documents received are referenced in ARC. This stage is essential, especially for the DSN, for which the number of files received over time is in the millions. Each DSN document corresponds to a declaration by a company (in the broad sense, now including public employers). This module therefore makes it possible to check that no declarations are duplicated or forgotten.

Next comes the **uploading**. The files are first read, then their data and structure defined in the XML[12] model of the DSN are stored in the database. It is during this stage that the files are associated with their "standard", which identifies the source and the edition, based on rules defined by the statistician. These two criteria, standard and edition, determine what will follow the processing.

Transformation into Tables that Statisticians Can Use

The DSN documents are XML files, the general and tree structures of which are documented in the DSN technical brief[13] very comprehensively. This format is adapted to the management software used by companies, but much less so than statistics. The purpose of the pipeline in ARC is to transform the DSN data into tables that can be used by statisticians.

Once integrated, the file data are structured in the **structuring** module. XML provides only hierarchical relationships between different data: for example, in the DSN, a company contains one or more establishments, on which one or more individuals depend, these being associated with one or more employment contracts. However, some management rules cannot be represented by simple hierarchical relationships, such as workplaces. Workplaces are not hierarchically dependent on the company, but there is a relationship between a contract of employment and the place where it is exercised. This relationship is formally documented in the DSN technical brief. The structuring stage therefore makes it possible to add around the data a set of management rules to create links between data and facilitate statistical operations.

[12] XML : eXtensible Markup Language (XML) est un langage utilisant des balises permettant de représenter des données de manière structurée.

[13] The "cahier technique de la DSN" (DSN technical brief) is a document describing the DSN exchange standard in detail: the meaning of each piece of data, value domains, controls, classification, structure of the messages transmitted, etc. *https://www.agirc-arrco.fr/mon-entreprise/specialistes-de-la-paie/declaration-sociale-nominative-dsn/* (Dubrulle et al., 2023).

> ***Statisticians may define compliance controls, form adjustments and filters for the data.***

The next stage is **control**. This module implements the processing stemming from the review and validation, and editing and imputation phases. Statisticians may define compliance controls, form adjustments and filters for the data. Thus, it is possible to control the format of a field, fully populate incomplete dates or filter by declaration dates.

The ***mapping*** module then formats the data in the model designed by the statistician for statistical use: this corresponds to the calculation of new variables and units.

The DSN files change from one edition to another. Statisticians can modify the rules used in the mapping to manage these changes, but independently of the statistical model, which can remain unchanged. For downstream applications, this consistency makes it possible to compare data from one year to the next without maintenance. The statistical model itself can evolve, but only marginally and in a manner controlled by the statistician.

This mapping stage requires the prior construction of the statistical model referred to in ARC as the “standard family”. This model makes it possible to relate statistical entities to each other and can either be defined directly in the application or imported from a DDI specification. The DSN model will contain tables, such as the Employer table, with fields containing the employer’s address, the company’s main activity or its registration number in the SIRENE[14] register (SIRET), or the Individual table with fields containing the individual’s forename or country of residence.

All these transformation modules can rely on external data integrated by statisticians, such as tables of nomenclatures, reference frames or correspondence tables. For example, for the DSN, the codes for employees’ countries of birth are recoded using the official geographical code. ARC thus partially performs the classification and coding phase.

Making the Data Files Available for Further Processing

On completion of the mapping, ARC finalises the data files. The DSN files, freshly transformed into usable data tables, are **made available** for **retrieval by client applications**. The application makes it possible to manage the clients of each data source. Each data delivery is time-stamped and only one data retrieval per client is authorised, so as to avoid duplicating data. Once the files have been downloaded by the declared clients, ARC deletes them after a period of time set by the statistician has elapsed.

The data made available undergoes other transformations by the ARC client applications: restructuring of data by statistical unit, calculation of derived variables, etc.

[14] SIRENE: Système Informatisé du RÉpertoire National des Entreprises et des établissements (National Enterprise and Establishment Register Database).

The integration of administrative data (Cotton and Haag, 2023) from the DSN goes through all these stages, from the receipt carried out by ARC to their transformation by client applications. The Répertoire Statistique des Individus et des Logements (Statistical Register of Individuals and Dwellings, RÉSIL[15]) uses the same integration scheme and ARC was chosen to perform the receipt function.

Receipt of the DSNs and the Desire for Reuse in INSEE's Information System

The ARC IT application was designed as part of the construction of a Système d'Information sur l'Emploi et les Revenus d'Activité (Employment and Activity Income Information System, SIERA), which receives inputs from various administrative data sources. It coordinates multiple applications and handles all processing operations, producing structural and short-term economic statistical indicators on employment and wages. More specifically, the objective of ARC within this information system was the receipt, from 2015 onwards, of the monthly Nominative Social Declaration files sent by the National Old Age Insurance Fund (Caisse nationale d'assurance vieillesse – Cnav[16]). This was a challenge for project management and the development team: the flow of data to be processed monthly was both massive and concentrated (originally, 1.8 million files per month, received between the 18th and 22nd of each month). File layouts were not stabilised and implementation times were short!

Two Fundamental Needs: Performance and Adaptability…

> ***ARC must be optimised constantly, in terms of both relevance and processing speed.***

In terms of functions, the product needed to meet two seemingly unrelated needs: to be powerful enough to process all the files each month within one week (the constraint as it was expressed at the outset) and yet to be able to adapt quickly to changes in those files. To that end, it must be optimised constantly, in terms of both relevance and processing speed. This requires flexibility and responsiveness in implementing changes related to the content or format of the source data, while minimising the impact on performance. The option decided upon involves leaving it to statisticians to program, test and re-program the processing operations for receipt in a "sandbox" environment according to changes to the sources and to the expectations of the downstream statistical processing. In addition, the application is designed such that the settings developed by statisticians are based only on configurations of the different phases of receipt, without affecting the processing, and therefore without any risk of altering the system's performance. Statisticians can thus focus on the "specialist profession" aspects and their colleagues in charge of digital use of the statistics can benefit from not needing to perform an optimisation phase.

[15] See the article by Olivier Lefebvre on RÉSIL in this issue.
[16] DSN data are produced by the Groupement d'intérêt public pour la modernisation des déclarations sociales (Public Interest Group for the Modernisation of Social Declarations, GIP MDS) *https://www.net-entreprises.fr/*.

… Which Make the Application More Sustainable?

The rapid development of this receipt function made it possible to design the ARC project in agile mode[17]. From a profession point of view, there was also a great need for such flexibility. This is because adaptive maintenance operations to adjust to changes in the standards of the source files made in other SIERA processing chains, such as the processing of the N4DS[18], were very costly. ARC needed to respond to this problem in a generic manner to ensure it could be reused and able to receive sources other than the DSN.

ARC was thus deployed in 2015 to receive the DSN XML files and then reused in SIERA to receive the files of the Déclarations Annuelles de Données Sociales (Annual declaration of social data, DADS), which had to "coexist" with the DSN until 2021. The application has also been used for the receipt of files already produced by INSEE, with a view to putting them in a shared format. In this initial release, the product already supported receipt of XML, CSV[19] and key-value[20] files.

The constraints that needed to be managed in order to develop a receipt tool for the DSN resulted in it being attributed with the "right properties" to designate it the central tool of a process for the receipt of administrative sources decoupled from downstream processing operations. The use of ARC then gradually expanded.

▶ First Uses Outside the Original Framework

Faced with a change in the standard for tax return files used in the ESANE process, the decision was made to use ARC rather than developing a new system specific to that process.

The completed technical appraisal highlighted several points: the need for the functional development of ARC, followed by the benefits of this tool for users and, finally, the fact that using the application represented the least expensive option to handle the functional development.

However, ARC's functional coverage was not entirely sufficient to handle complex hierarchical files such as the tax files. Nevertheless, the standardisation of the receipt process offered by the application has made it possible to carry out the changes quickly and to enrich the product's offering. ARC was therefore reused outside of SIERA for the first time in 2019 for the receipt of tax files in ESANE.

[17] The purpose of the agility is to direct efforts towards what has the most value for the user, by adapting to changes at the lowest cost.

[18] N4DS: standard for Déclarations Dématérialisées Des Données Sociales (Digitised Social Data Declarations), used by DADS.

[19] CSV is the name of a file format intended to present data separated by commas. This is a simplified means of displaying data so that such data can be transferred between programs.

[20] The key-value storage format matches keys (for example, profession headings) with values.

In 2020, as part of a working group of the European Statistical System on the sharing of statistical tools, its functions were again extended in two directions. The first of these enabled the application to be deployed on containerised infrastructures[21], allowing increased scalability[22] in particular. The second made it possible to use the file receipt stages in web-service mode. ARC then became capable of processing unit requests on demand, in addition to the mass processing initially developed. SIRENE4 has thus integrated the application into this machine-to-machine mode of use to automatically verify conformity of the registration files from the Guichet Unique[23] (One-Stop Shop) (Alviset, 2020).

ARC has thus evolved gradually over time to become a robust and efficient application (***Box***).

▶ Box. Technical Specifications and Performance.

ARC is an open-source ETL (*Extract Transform Load*) developed by INSEE. The source code of the application is hosted on the inseeFr section on GitHub: *https://github.com/InseeFr/ARC*

ARC offers a web module, a batch module and a web-service module. Each module is containerised with Docker* and can be deployed autonomously in accordance with business needs. The containers are available on *dockerhub*: *https://hub.docker.com/u/inseefr*

The web module offers a human-machine interface to configure and launch processing operations on files and potentially control mass batch processing. The batch module makes it possible to process massive file flows and ensure recovery should an error occur. The web-service module offers a data recovery service and a unit file processing service.

ARC can process XML, key-value and text files in CSV, delimited or positional formats.

ARC's performance has improved considerably since its creation: the first monthly DSN uploads in 2015 took 9 days, compared to 60 hours today, even though the volume of data to be processed is twice as large. This significant improvement was made possible by the decoupling of data storage and processing, in particular.

Processing is performed in ARC using PostgreSQL databases. The iteration dedicated to uploading the DSN currently has a single database with 32 CPUs and 32 GB of RAM.

The latest version of ARC features horizontal scalability. Rather than having a single database with many resources, the application can use several small databases in tandem. This architecture makes it possible to avoid the processing capacity problems inherent in the use of a single machine: the processing time decreases in proportion with the number of databases dedicated to the application.

* *Docker is a system allowing the creation, sharing and execution of containers.*

▶ Change of Scale: the Organisation of Pooling

The gradual expansion of the functions and uses of ARC, together with its indispensable nature in various production operations, have led to an approach that is adapted to these issues. The aim was to manage the implementation and deployment of cross-functional investments (performance improvement, metadata processing, maintenance of operational and security conditions, etc.) and to generate engagement within the user community (communications regarding the product and its developments, collection and

[21] A container infrastructure makes it possible to automate container deployment, scaling and management. A container is an execution environment that contains all the necessary components (code, dependencies and libraries) to execute the application code without using the dependencies of the host machine.

[22] Ability to adapt to significant changes in the volume of data to be processed.

[23] The Guichet électronique des formalités d'entreprises (Electronic Business Formalities Window) (Guichet Unique - One-Stop Shop) is a secure Internet portal, to which every company has been required to report its creation, as well as various significant events, since 1 January 2023.

prioritisation of needs, training, etc.). These actions are essential for a pooled product such as ARC in order to maintain functional coverage and take into account the needs of the various users.

In 2021, the RÉSIL programme became the project owner of the ARC product, not only due to its central position within the information system, but also because of the range of sources it receives.

▶ Conclusion

The world of data, like that of computer science, is constantly evolving. To address the challenges they face in relation to information technology, companies have developed strategies such as implementing agility and DevOps[24]. This idea was taken up by the world of data, with the development of DataOps[25], which aims in particular to reconcile automation, reproducibility, interactivity and traceability for data processing, while bringing together different data professions around common tools. By offering agility in its developments, INSEE already incorporates many DataOps ideas. The receipt service provided by ARC is fully in line with this approach, thanks to the flexibility and decoupling of the processing that it implements.

ARC is an application that has evolved over time in accordance with a range of varied needs. It is an application that is now easy to deploy and is suitable for flexible and efficient loading of external data, upstream of statistical applications. It is becoming central to the implementation of the strategy of using administrative data that encompasses the diversification of sources, the speed of their processing and the ability to make them available for use in a variety of statistical processes, all in complete security.

Opening it up for more exploratory uses, carried out independently by statisticians, can make it easier to explore the potential of new data sources, with a view to statistical innovation.

[24] DevOps is a movement in computer engineering and a technical practice aimed at unifying software development and IT infrastructure administration.

[25] DataOps is an automated method for improving quality and reducing the time required for data analysis. *https://dataopsmanifesto.org/en/*.

Bibliography

- ALVISET, Christophe, 2020. La troisième refonte du répertoire Sirene : trop ambitieuse ou pas assez ? In: *Courrier des statistiques*. [online]. 29 June 2020. Insee. N°N4, pp 101-121. [Accessed 16 May 2024]. Available at: *https://www.insee.fr/fr/information/4497083?sommaire=4497095*.

- BISEUL, Xavier, 2023. Data lineage : définition, principes et outils. In: *Journal du Net*. [online]. 28 February 2023. [Accessed 14 March 2024]. Available at: *https://www.journaldunet. fr/web-tech/guide-du-big-data/1516833-data-lineage-definition-principes-et-outils/*.

- BONNANS, Dominique, 2019. RMéS: INSEE's Statistical Metadata Repository. In: *Courrier des statistiques*. [online]. 27 June 2019. Insee. N° N2, pp. 46-57. [Accessed 14 March 2024]. Available at: *https://www.insee.fr/en/information/4195079?sommaire=4195125*.

- COTTON, Franck and DUBOIS, Thomas, 2019. Pogues, a Questionnaire Design Tool. In: *Courrier des statistiques*. [online]. 19 December 2019. Insee. N° N3, pp. 17-28. [Accessed 14 March 2024]. Available at: *https://www.insee.fr/en/information/5014167?sommaire=5014796*.

- COTTON, Franck and HAAG, Olivier, 2023. Integrating administrative data into a statistical process–Industrialising a key phase. In: *Courrier des statistiques*. [online]. 30 June 2023. Insee. N° N9, pp. 104-125. [Accessed 14 March 2024]. Available at: *https://www.insee.fr/en/information/8232639?sommaire=8232646*.

- DUBRULLE, Bertrand, ROSEC, Olivier and SUREAU, Christian, 2023. Une norme d'échange pour alimenter des référentiels et en assurer la qualité. In: *Courrier des statistiques*. [online]. 30 June 2023. Insee. N°N9, pp 126-146. [Accessed 18 June 2024]. Available at: *https://www.insee.fr/fr/information/7635835?sommaire=7635842*.

- DONDON, Alexis and LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In: *Courrier des statistiques*. [online]. 30 June 2023. Insee. N°N9, pp 86-103. [Accessed 14 March 2024]. Available at: *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.

- ERIKSON, Johan, 2020. Using a Process Model at Statistics Sweden Implementation, Experiences and Lessons Learned. In: *Courrier des statistiques*. [online]. 29 June 2020. Insee. N° N4, pp. 122-141. [Accessed 14 March 2024]. Available at: *https://www.insee.fr/en/information/6050996?sommaire=6049874*.

- HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative. Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. In: *Courrier des statistiques*. [online]. 6 December 2018. Insee. N° N1, pp. 25-34. [Accessed 14 March 2024]. Available at: *https://www.insee.fr/fr/information/3647025?sommaire=3647035*.

- LAMARCHE, Pierre and LOLLIVIER Stéfan 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In: *Courrier des statistiques*. [online]. 8 July 2021. Insee. N°N6, pp 28-46. [Accessed 31 May 2024]. Available at: *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- RENNE, Catherine, 2018. Understanding the Nominative Social Declaration (DSN) for Better Statistical Measurement. In: *Courrier des statistiques*. [online]. 6 December 2018. Insee. N° N1, pp. 35-44. [Accessed 14 March 2024]. Available at: *https://www.insee.fr/en/information/4195367?sommaire=4195376*.