L'accueil des données administratives : un processus structurant

Olivier Lefebvre[*], *Manuel Soulier*[**] *et Thomas Tortosa*[***]

Compte tenu de l'importance croissante des données administratives dans le processus d'élaboration des statistiques, rationaliser leur traitement devient un enjeu essentiel.
Et cela dès l'accueil de ces données ! On les reçoit telles que configurées en fonction de leurs usages administratifs et il faut les transformer en données statistiques, c'est-à-dire en données organisées selon les unités statistiques d'intérêt (individus, ménages, établissements, employeurs, etc.) et les concepts statistiques. Cette phase peut être mutualisée et « découplée » des processus statistiques en aval, offrant ainsi davantage de possibilités d'adaptation, mais aussi de partage d'information. L'enjeu est de construire un dispositif alliant adaptabilité, performance, sécurité et traçabilité.
L'Insee a engagé cette démarche avec l'outil ARC (Accueil-Réception-Contrôle) à partir d'un cas d'usage exigeant : le traitement mensuel d'environ deux millions et demi de fichiers de déclarations sociales nominatives (DSN). En étendant progressivement ses fonctionnalités et ses performances pour l'adapter à de nouvelles données et à de nouvelles contraintes, ARC constitue désormais un composant essentiel du dispositif de production statistique de l'Insee.

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.

* Maître d'ouvrage du programme Résil, DSDS, Insee.
olivier.lefebvre@insee.fr

** Chef de projet Informatique sur l'application ARC (Accueil Réception Contrôle), Direction régionale du Centre – Val de Loire, Insee.
manuel.soulier@insee.fr

*** Chef de projet statistique pour le projet « accueil des sources » au sein du programme Résil, DSDS, Insee.
thomas.tortosa@insee.fr

Industrialiser l'intégration des données administratives dans nos systèmes d'information est essentiel, compte tenu de l'importance croissante de ce type de données dans nos processus de production statistique (Cotton et Haag, 2023). Pour relever ce défi, la solution adoptée par l'Insee est une structure d'accueil des sources offerte aux producteurs de statistiques, reposant sur un outil générique moderne et mutualisé.

Au début de l'utilisation de données administratives, ce travail d'intégration se faisait isolément d'une source à l'autre : chacun développait son propre processus, adapté à la source traitée et aux traitements en aval. Ce modèle a fonctionné pendant des décennies. Cependant, dans le courant des années 2010, ces données administratives sont devenues plus nombreuses, plus fréquentes, plus évolutives, et de surcroît susceptibles d'alimenter plusieurs chaînes de production de données. Par ailleurs, des besoins nouveaux ont émergé dans le processus d'accueil qui devait être plus « adaptable », performant, traçable, ouvert, tout en continuant d'assurer un niveau élevé de sécurité.

Les qualités attendues d'un service d'accueil des fichiers administratifs dans un univers statistique : adaptabilité, performance, traçabilité et sécurité

Le service doit être **adaptable**. Il doit prendre en compte au plus vite des changements de contenu ou de format de l'information transmise. De telles transformations sont inévitables, car à l'image des politiques publiques ou de leurs processus de mise en œuvre, la donnée administrative n'est pas figée et évolue en fonction des impératifs de la politique publique qu'elle accompagne.

Ces évolutions doivent pouvoir être appliquées rapidement, sans pour autant être propagées simultanément à tous les fichiers si ceux-ci proviennent d'organismes différents. Les fichiers à accueillir peuvent coexister dans plusieurs versions avec des contenus modifiés selon la date de conception.

Pour répondre à ces besoins d'adaptabilité et de réactivité, le système doit d'une part gérer et accueillir simultanément plusieurs versions de fichiers, et d'autre part proposer des traitements dits « génériques », c'est-à-dire fonctionnant sur tous les fichiers.

> ***Les statisticiens doivent disposer dans la fonction d'accueil, d'outils pour paramétrer les traitements et mesurer l'impact des changements de paramètres sur les données.***

Aussi, les statisticiens doivent disposer dans la fonction d'accueil, d'outils pour paramétrer les traitements et mesurer l'impact des changements de paramètres sur les données.

En outre, cette fonction doit tout à la fois accueillir des fichiers dont le format et le contenu peuvent être amenés à évoluer rapidement et alimenter des chaînes statistiques utilisatrices de données stables sur le long terme. Étendre le principe de généricité de l'outil servant d'interface entre l'accueil et les applications clientes est une solution pour répondre à cette problématique.

La **performance** informatique devient cruciale. En effet, les données sont transmises de plus en plus souvent et doivent être valorisées rapidement. Désormais, des sources extrêmement lourdes sont reçues chaque mois par l'Insee et nécessitent d'être traitées rapidement pour assurer la pertinence des statistiques produites dans des délais toujours plus réduits (demande sociale européenne).

La **traçabilité** est une exigence de qualité fondamentale dans un processus de production. Plus les changements sont nombreux, plus il est indispensable d'en conserver la trace. Cela permet, si nécessaire, de reproduire le traitement, de rendre compte des opérations réalisées et ainsi d'analyser plus facilement les évolutions à mener sur les traitements en aval, et enfin de documenter le processus.

Les données administratives constituent une source importante de données. Elles alimentent depuis longtemps des processus de production mais pourraient servir à des tests de nouveaux traitements statistiques. Avec les nouveaux métiers de la science des données et la montée en puissance à l'Insee des métiers de *data scientists*, il est nécessaire de prévoir des mécanismes permettant d'ouvrir, de manière contrôlée, l'accès aux données brutes statistiques[1] afin de les exploiter de façon innovante.

Enfin, la **sécurité** s'avère primordiale pour ce type de données, celles-ci pouvant contenir des données personnelles imposant une stricte confidentialité, tout comme la protection de leur intégrité ; l'exigence de traçabilité évoquée plus haut participe également de la sécurité d'ensemble du processus et des données traitées.

Prévoir un accueil transverse des données implique donc de centraliser les règles de sécurité dans le cadre d'une gouvernance des données administratives. Il s'agit par exemple de mutualiser et réaliser le plus tôt possible des processus transversaux tels que la « pseudonymisation » des données (Cotton et Haag, 2023), mettre en place une politique et des outils de gestion des droits d'accès, tout en laissant la possibilité à chaque propriétaire d'appliquer des règles spécifiques supplémentaires.

> ***La mutualisation d'un outil d'accueil de données administratives s'est imposée.***

Ces exigences nécessitent des investissements significatifs ; ainsi s'est imposée la mutualisation d'un outil d'accueil de données administratives, capable de gérer des sources de natures et d'origines différentes, et d'alimenter des processus d'exploitation divers. Un tel outil repose sur un découplage de la fonction d'accueil de la donnée et de la fonction de traitement ou d'analyse de celle-ci.

1 Les données brutes statistiques sont des données administratives existant sous un format exploitable statistiquement. Elles sont « brutes » du point de vue du statisticien, car elles n'ont pas encore été traitées pour un usage statistique.

▶ Découpler la phase d'accueil des données de celle des traitements statistiques…

Constituer le service d'accueil des données, c'est penser la phase d'accueil comme une activité à part entière, qu'il faut dissocier des traitements nécessaires à la création d'un produit statistique.

Traditionnellement, le processus d'élaboration d'un produit statistique s'appuyant sur des données externes est de type itératif. Après une étape d'appropriation, le statisticien intègre son fichier pour obtenir un résultat attendu via différentes étapes. Si ces dernières peuvent varier, on retrouve les suivantes :

- si besoin, structurer et transformer le fichier en base de données ;
- renommer les variables afin de pérenniser le traitement ou expliciter le nom de celles-ci ;
- retraiter les variables afin de corriger certaines imperfections (non-réponse, valeurs aberrantes, etc.) ;
- créer des variables statistiques issues d'une ou plusieurs variables parfois modifiées ou agrégées (toutes celles constituant le salaire, tous les revenus d'une catégorie d'agents) ;
- réaliser un produit de diffusion.

Ces phases sont la plupart du temps implémentées par blocs.

▶ … mutualiser l'accueil des données…

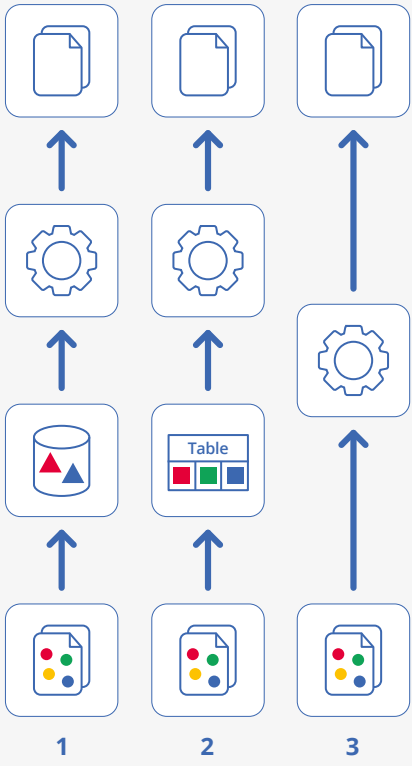
Le statisticien dispose de l'intégralité des données contenues dans le fichier et, par étapes, il va le transformer en un produit statistique. Toutefois, s'il est confortable lorsqu'on le crée, ce processus devient rapidement difficile à maintenir et à exploiter s'il n'est pas développé et implémenté de façon modulaire. En cas de transformations des données, modifier la chaîne de production peut vite s'avérer complexe, si la phase de traitement est adhérente à la phase d'accueil, c'est-à-dire si les traitements s'appuient directement sur les données brutes administratives.

L'idée de découpler l'accueil des données s'est progressivement imposée (***figure 1***). Ainsi, l'Insee a mis en place l'accueil de la Déclaration Sociale Nominative (DSN)[2], et développé un outil dédié (Accueil-Réception-Contrôle, dit ARC), pour alimenter la production de statistiques sur l'emploi et les revenus d'activité (Renne, 2018). C'est également le cas dans le processus Esane (Élaboration de Statistiques Annuelles d'Entreprises) qui s'appuie d'une part sur des données d'enquêtes et d'autre part sur des données fiscales. Plus récemment, le projet de refonte de Fidéli[3] à la suite de la disparition de la taxe d'habitation, a partagé l'application en deux : la première

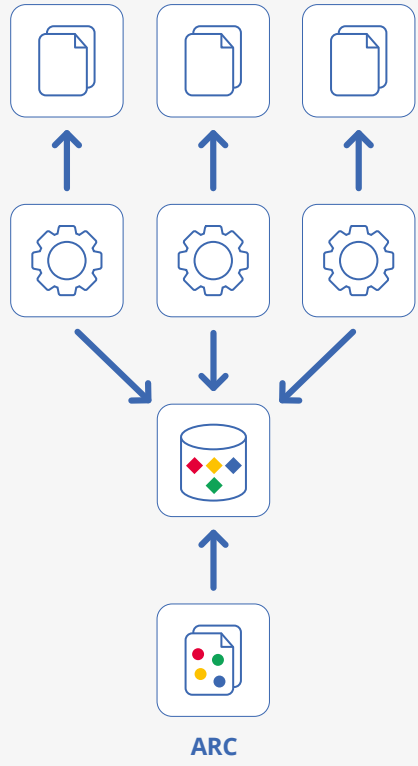
2 La Déclaration Sociale Nominative (DSN) est une déclaration obligatoire, unifiée et en ligne, établie par chaque employeur. La DSN permet d'assurer le recouvrement des cotisations et d'attribuer des droits aux salariés.
3 Fidéli : Fichier Démographique d'origine fiscale sur les Logements et les Individus (Lamarche et Lollivier, 2021).

pour assurer l'accueil des données et la seconde pour effectuer les traitements. Réalisé ex-post au regard des expériences, séparer l'accueil de sources externes des traitements statistiques est un choix d'une utilité majeure pour les statisticiens.

Figure 1 - L'accueil, première étape de la rationalisation des traitements



Lecture : Les données sont accueillies plusieurs fois. Chaque pipeline gère différemment les données en entrée. Le pipeline 1 sélectionne les données utiles et modifie la forme mais aussi le contenu, qu'il stocke en base de données. Le pipeline 2 sélectionne les données et les affine avant de les stocker dans une table exploitable par un logiciel statistique (SAS ou R, par exemple). Quant au pipeline 3, il intègre en bloc le fichier dans ces traitements.



Lecture : Le service d'accueil ARC restructure l'information du fichier, mais sans la modifier. Chaque pipeline peut interroger le fichier pour récupérer les données.

▶ … Pour un usage rationalisé et maîtrisé !

Avec ce découplage, le système d'information est alors plus robuste. L'accueil des sources absorbe en grande partie les chocs exogènes, si ce service a été conçu pour cela[4]. Enlever, ajouter ou modifier des informations peut être traité dans cette phase afin d'alimenter les traitements statistiques en aval de façon quasi identique, en minimisant la maintenance de ce dernier. ARC repose sur ce principe : une phase de conceptualisation des données, c'est-à-dire de transformation des données reçues en un système de données brutes statistiquement exploitables. Cette phase permet de gérer des modifications, comme renommer des variables ou modifier leur contenu lorsque cela est possible.

Par exemple, le fichier des déclarations de revenus POTE[5] de la DGFiP[6] est au format texte ; la lecture n'est donc possible que grâce à un dessin de fichier.

Position		Lg	Numéri-coualpha	Format de lecture	Format d'écriture	Input format	NomPAC	Libellé
247	254	8	9(8)	8.		8.	DADOKZ	SITFAM : DATE DECES DE LA 2042

Ici la variable DADOKZ, qui correspond à la date de décès du référent fiscal du foyer, se lit dans le fichier de la position 247 à 254 (longueur 8). Le format indiqué est un format numérique (longueur 9), alors que la nature de la variable est de type date.

À l'issue du processus d'accueil, la donnée sera mise à disposition dans une variable date_dc (qui explicite le contenu de la variable) au format date (« YYYY-MM-DD »).

Le découplage rend possible et simplifie l'utilisation multiple des données. Lorsque l'accueil des données est intégré dans une chaîne de production d'une application de production statistique, il est difficile, voire impossible, d'ouvrir cet accès aux données à d'autres applications.

Par exemple, les données des déclarations sociales nominatives (DSN) chargées dans ARC étaient initialement destinées à la chaîne structurelle du calcul de l'emploi salarié.

Cette dernière produit les statistiques annuelles de l'emploi en matière de stock ou de répartition par statut ou activité économique. Pour cette utilisation, la chaîne structurelle est « cliente » et consommatrice des données DSN accueillies dans ARC.

Lorsque l'Insee a souhaité utiliser ces mêmes données pour réaliser des estimations conjoncturelles de l'emploi salarié, l'adaptation a été simplifiée par le découplage : à l'instar de la chaîne structurelle, la nouvelle chaîne conjoncturelle a été déclarée « cliente » de ARC. Cela aurait été quasi impossible si l'accueil de la DSN avait été intégré et couplé à la chaîne structurelle.

4 Voir ci-après la partie sur les nouveaux métiers.
5 Fichier « Permanent des Occurences de Traitement des Émissions » : il contient les données relatives aux déclarations des revenus de l'année transmises par les contribuables à la DGFiP au printemps de l'année suivante.
6 DGFiP : la direction générale des Finances publiques est une direction de l'administration publique centrale française qui dépend du ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique.

> ***Isoler le processus d'accueil donne l'opportunité de gérer les droits d'accès aux données par les applications clientes.***

De plus, isoler le processus d'accueil donne l'opportunité de gérer les droits d'accès aux données par les applications clientes, chacune d'elles pouvant sélectionner les données dont elle a besoin parmi celles mises à disposition. Le partage se fait à la source sans avoir besoin de construire une « passerelle » entre les applications.

Enfin, ce découplage permet aussi d'identifier l'accueil des données comme un processus à part entière, avec tous ses avantages. Il est ainsi possible de le découpler des processus clients et donc de le faire évoluer de façon indépendante, ou encore d'ajouter un processus client supplémentaire alimenté à la source. Cela permet aussi de « cibler » les réflexions et investissements pour l'optimiser.

▶ Gérer et « activer » les métadonnées

> ***Sans métadonnées, les données mises à disposition sont inexploitables.***

Sans métadonnées, les données mises à disposition sont inexploitables : en effet, elles sont issues d'une phase de transformation des données administratives en concepts statistiques. Les métadonnées servent à documenter les données produites et sont dans ce cas dites « passives ». Elles sont générées au fil des traitements et permettent de garder en mémoire les opérations sur les données, ainsi que l'utilisation des variables en interne comme en externe. Pour cela, l'Insee dispose d'un référentiel de métadonnées statistiques, RMéS, qui en permet la gestion, le partage et la diffusion (Bonnans, 2019).

Au sein du domaine des enquêtes, les métadonnées sont exploitées en entrée du processus de conception du support de collecte afin de le générer (Cotton et Dubois, 2019). Cette utilisation est possible grâce à la mise en place d'un ensemble d'outils et de services reliés à RMéS. On parle alors de métadonnées « actives », c'est-à-dire qu'elles ont une fonction autre que documentaire dans le processus statistique. Un des enjeux de l'accueil des données administratives est donc de rendre ces méta-données « activables », à l'instar de ce qui se pratique sur les enquêtes ; cela implique de documenter le plus en amont possible les métadonnées issues des fichiers administratifs.

Fournir aux producteurs les métadonnées associées aux données statistiques brutes leur donne les éléments pour documenter leur processus et, in fine, intégrer leurs propres métadonnées au sein de RMèS. Une expérimentation récente sur des données foncières a démontré que la majeure partie des métadonnées saisies dès la phase d'accueil pouvaient être réutilisées sans modifications dans les processus ultérieurs, et ce jusqu'à la réalisation des bases de diffusion.

Par ailleurs, livrer des métadonnées en entrée du processus de traitement permet de mettre en place des métadonnées dites de production : par opposition aux métadonnées dites de diffusion, elles permettent de tracer, configurer voire spécifier les modifications successives des données. On parle alors de lignage des données[7] (Biseul, 2023).

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

Mettre en place un service d'accueil exige de spécialiser cette phase d'accueil des données. Qui dit spécialisation, dit métiers. Au regard des différentes tâches à réaliser, deux nouveaux métiers ont émergé.

Le premier est celui de **modélisateur des données**. Celui-ci conçoit les modèles dans lesquels seront insérées les données administratives pour les utilisateurs. Il transforme un modèle donné, sur lequel il ne peut agir, en un modèle statistiquement exploitable ; ce dernier doit être à la fois robuste aux changements et construit de telle sorte que les utilisateurs puissent utiliser simplement les données pour produire des statistiques. Il lui faut donc être compétent en matière de modélisation mais aussi être à l'écoute des utilisateurs afin que le modèle développé corresponde aux attentes.

Les données étant par nature administratives, il est parfois difficile de les transformer en concept statistique. Le modélisateur peut les segmenter par thématique : on parle alors de partitionnement vertical des données. Une modélisation technique vient compléter la modélisation sémantique.

Par exemple, le fichier des déclarations de revenus (POTE), fourni par la DGFiP, est un fichier très volumineux, tant par son nombre de lignes (45 millions) que par son nombre de variables (600 pour la partie fixe, plus du double pour la partie variable). Dans ce fichier, il est possible d'identifier plusieurs thématiques autour de l'impôt : impôt sur le revenu, contribution sociale généralisée, impôt sur la fortune immobilière, etc. Autant de « thèmes » qui peuvent être isolés. L'existence et les caractéristiques d'un impôt ne sont par nature pas pérennes : la taxe d'habitation en est un exemple. L'avantage de cette modélisation est de construire un modèle autour de thématiques pour chaque impôt. In fine, l'utilisation du fichier sera plus robuste puisque seules les parties affectées par les changements sont modifiées.

Le second métier nouveau est celui d'**intendant des données**. L'intendance des données est un concept pour lequel il existe plusieurs définitions, correspondant à des contours plus ou moins ambitieux. La définition retenue dans cet article est celle de Statistique Canada[8], à savoir : « L'intendance des données, c'est la gouvernance des données en action, c'est-à-dire la déclinaison opérationnelle de la politique en matière de données ». Il s'agit donc de la mise en œuvre effective des règles régissant la collecte, la gestion, la sécurité, la qualité et la diffusion des données au sein d'une organisation.

7 *https://www.journaldunet.fr/web-tech/guide-du-big-data/1516833-data-lineage-definition-principes-et-outils/.*
8 *https://www.statcan.gc.ca/fr/afc/litteratie-donnees/catalogue/892000062020013.*

Ainsi, pour l'intendant, sont requises des compétences en matière d'administration des données puisqu'il assure la réception, le contrôle et la documentation des données qu'il met à disposition.

Il doit aussi posséder des compétences en matière de gestion de données, puisqu'il assure la gestion des droits d'accès aux données et suit les conventions passées avec les fournisseurs (respect des délais et du format de transmission et de conservation des données).

Concernant l'accès aux données, la diversité des utilisations possibles pour une même donnée administrative complexifie la problématique de leur sécurisation. « Pseudonymisées » et transformées en un format statistique, les données sont plus ouvertes et plus « partageables ». Cependant, ce partage doit être organisé et sélectif, selon des finalités précises et en conformité avec le principe de proportionnalité des traitements. L'intendant des données doit appliquer le droit des utilisateurs concernant leur accès aux données.

Enfin, ce spécialiste est en contact direct avec les utilisateurs et les producteurs de données administratives. Il est notamment en première ligne pour régler les problèmes de transmission de données. Des qualités relationnelles sont donc également requises.

L'intendant est légitime pour la collecte de la donnée et sa gestion mais pas sur ses usages. Il faut donc conserver une relation entre le producteur statistique et le fournisseur, apurée des questions de gestion, donc centrée sur le contenu et l'usage des données ainsi que sur leurs évolutions.

▶ Accueil – Réception – Contrôle (ARC), le service informatique d'accueil des données de l'Insee —

L'application informatique Accueil-Réception-Contrôle ARC assure le service d'accueil à l'Insee depuis une dizaine d'années, puisqu'elle existe depuis 2015 avec l'accueil des fichiers de la DSN.

> ***L'application informatique Accueil-Réception-Contrôle ARC assure le service d'accueil à l'Insee depuis une dizaine d'années.***

ARC couvre fonctionnellement certaines des phases du GSBPM[9] (***figure 2***). Ce dernier est un cadre standard des organismes statistiques qui leur permet d'adopter une terminologie commune pour décrire le cycle de vie d'une opération statistique (Erikson, 2020).

9 Le modèle générique de description des processus de production statistique (GSBPM pour *Generic Statistical Business Process Model*) décrit les différentes étapes à suivre pour produire des statistiques publiques.

Figure 2 - Le champ fonctionnel d'ARC au regard du modèle générique de description des processus de production statistique (GSBPM)

Légende :

- Ces phases sont des « prérequis » au bon fonctionnement d'ARC, à gérer en amont et en dehors.
- Ces phases sont couvertes par ARC sans que le statisticien puisse les paramétrer.
- Ces phases sont celles que le statisticien peut paramétrer dans ARC.
- Phase en cours de développement.


Gestion de la qualité / Gestion des métadonnées
Traitement
Analyse
Diffusion
Évaluation
5.1 Intégration des données
5.2 Classification et codage
5.3 Examen et validation
5.4 Édition et imputation des données
5.5 Calcul de nouvelles variables et unités
5.6 Calcul des coefficients de pondération
5.7 Calcul des agrégats
5.8 Finalisation des fichiers de données
6.1 Élaboration du projet de produits
6.2 Validation des produits
6.3 Interprétation et explication des produits
6.4 Mise en place du contrôle de la divulgation
6.5 Finalisation des produits
7.1 Actualisation des systèmes de produits
7.2 Élaboration des produits de diffusion
7.3 Gestion de la publication des produits de diffusion
7.4 Promotion des produits de diffusion
7.5 Gestion de l'assistance aux utilisateurs
8.1 Recueil des produits d'évaluation
8.2 Conduite de l'évaluation
8.3 Adoption d'un plan d'action

Le statisticien s'appuie sur l'opération transverse de gestion des métadonnées pour élaborer la conception des produits finaux de diffusion et la conception de la description des variables utilisées. Pour cela, il utilise directement l'implémentation proposée dans l'application ARC ou lui soumet une modélisation DDI[10] (Dondon et Lamarche, 2023) réalisée via l'outil Colectica Designer[11].

Les phases de conception de la collecte et de conception du cadre et de l'échantillon sont hors du champ de l'application ARC et sont élaborées en accord avec le fournisseur de données.

Les fonctionnalités couvrant les phases de conception du traitement et de l'analyse, de conception de système de production et du déroulement des travaux et les trois phases d'élaboration modélisent la fonction d'accueil dans ARC et posent le cadre des configurations possibles pour le statisticien. Elles ont été construites lors de la conception de l'application et constituent le pipeline d'ARC (***figure 3***).

Le statisticien a toutefois la main pour configurer le déroulement de travaux pour les phases de l'étape « Traitement » ; il peut tester ses configurations sur de petits volumes de données, dans des espaces dédiés, disjoints des espaces de production et appelés « bacs à sable ». Les bacs à sable dans ARC permettent la mise à l'essai du système de production sur un nombre réduit de fichiers sources. Lorsque la mise au point est terminée, le statisticien applique ses configurations sur le flux de fichiers réels et procède à la finalisation du système de production.

Dans le cadre d'un traitement statistique utilisant une fonction d'accueil, ARC couvre les étapes « conception », « élaboration » et « collecte » du GSBPM et partage avec les applications utilisatrices des données, certaines phases de l'étape « Traitement ». Par exemple, ARC intervient sur la phase d'édition et imputation des données[12] pour publier certaines données afin de les rendre exploitables statistiquement (correction de modalités, mise en conformité au format ou aux valeurs attendues), alors que les transformations statistiques au sens métier (imputations de valeurs manquantes ou détection de valeurs aberrantes) sont déportées sur les applications de traitement des données.

Il n'y a pas à ce jour de production automatisée d'un bilan qualité de l'accueil d'une source (nombre d'enregistrements lus, nombre de valeurs erronées, etc.), qui correspond à la phase de recueil des produits d'évaluation du GSBPM. Celle-ci est actuellement prise en charge par le système d'information (SI) client. Le produit ARC devra évoluer afin de proposer une implémentation de ce recueil pour les traitements le concernant (et notamment le respect des normes annoncées), car l'assurance qualité est une composante essentielle de la fonction d'accueil.

10 DDI : *Data Documentation Initiative* est un consortium international d'instituts de recherche et de producteurs de statistiques qui vise à définir des standards pour la documentation des données statistiques, avec un focus particulier sur les données d'enquêtes, des méthodes de collecte et des référentiels (nomenclatures, codifications, etc.) utilisés pour la collecte. Le format DDI repose sur le format XML (voir définition du format XML plus loin).

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

12 « Data editing and imputation » en anglais.

▶ Figure 3 - Le pipeline de la Déclaration Sociale Nominative (DSN)

▶ Le pipeline de traitement à travers l'exemple de la DSN —

> ***Les déclarations sociales nominatives sont converties en données statistiques par le traitement informatique ARC, via une succession de modules fonctionnels. Chaque module réalise une opération précise.***

La DSN est une déclaration en ligne obligatoire, permettant aux employeurs de transmettre aux organismes de protection sociale les informations relatives aux salariés (Humbert-Bottin, 2018). En décembre 2023, l'Insee a reçu environ 2,5 millions de fichiers d'employeurs avec les données salariales de leurs employés. Il en recevait 1,8 million en 2016.

Les DSN sont converties en données statistiques par le traitement informatique ARC. Ce traitement est architecturé en « pipeline », c'est-à-dire constitué d'une succession de modules fonctionnels, chaque module réalisant une opération précise. Les modules de ARC sont préalablement paramétrés par le statisticien. ARC suit une logique de document : chaque fichier est traité individuellement et indépendamment des autres, et son nom est conservé comme identifiant pour chaque donnée.

Intégrer les données administratives

La phase GSBPM d'intégration des données est couverte dans ARC par les deux premiers modules du pipeline de traitement.

Dans le module **réception**, les documents reçus sont référencés dans ARC. Cette étape est essentielle, d'autant plus pour la DSN pour laquelle le nombre de fichiers reçus au fil de l'eau se compte en millions. Chaque document de la DSN correspond à la déclaration d'une entreprise (au sens large, y compris les employeurs publics désormais). Ce module permet donc de vérifier qu'aucune déclaration n'est en double ou oubliée.

Puis vient le **chargement**. Les fichiers sont d'abord lus, puis leurs données et leur structure définie dans le modèle XML[13] de la DSN sont stockées dans la base de données. C'est lors de cette étape que les fichiers sont associés à leur « norme », laquelle identifie la source et le millésime, selon des règles définies par le statisticien. Ces deux critères de norme et de millésime déterminent quelle sera la suite du traitement.

Les transformer en tables utilisables par le statisticien

Les documents de la DSN sont des fichiers XML dont la structure et l'arborescence sont documentées dans le cahier technique de la DSN[14], de manière très complète. Ce format est adapté aux logiciels de gestion utilisés par les entreprises, mais beaucoup moins à

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

14 Le « cahier technique de la DSN » est un document décrivant de manière détaillée la norme d'échange de la DSN : signification de chaque donnée, domaines de valeurs, contrôles,nomenclatures, structure des messages transmis, etc. *https://www.agirc-arrco.fr/mon-entreprise/specialistes-de-la-paie/declaration-sociale-nominative-dsn/* (Dubrulle et alii, 2023).

la statistique. Le but du pipeline dans ARC est de transformer les données de la DSN en tables exploitables par les statisticiens.

Une fois intégrées, les données des fichiers sont structurées dans le module de **structuration**. Le XML ne fournit que des relations d'ordre hiérarchique entre les différentes données : par exemple, dans la DSN, une entreprise contient un ou plusieurs établissements, desquels dépendent un ou plusieurs individus, qui sont associés à un ou plusieurs contrats de travail. Or certaines règles de gestion ne peuvent pas être représentées par simple relation hiérarchique, comme les lieux de travail. Ces derniers ne dépendent pas hiérarchiquement de l'entreprise, mais il existe une relation entre un contrat de travail et le lieu où il est exercé. Cette relation est documentée formellement dans le cahier technique de la DSN. L'étape de structuration permet donc d'ajouter autour des données, un ensemble de règles de gestion pour créer des liens entre données et faciliter les opérations statistiques.

> ***Le statisticien peut définir des contrôles de conformité, des redressements de forme, des filtres sur les données.***

L'étape suivante est le **contrôle**. Ce module implémente les traitements relevant des phases d'examen et validation et d'édition et imputation. Le statisticien peut définir des contrôles de conformité, des redressements de forme, des filtres sur les données. Ainsi, il est possible de contrôler le format d'un champ, compléter des dates non renseignées, ou filtrer sur les dates de déclaration.

Le module de ***mapping*** vient ensuite mettre au format les données dans le modèle conçu par le statisticien pour l'exploitation statistique : cela correspond au calcul de nouvelles variables et unités.

Les fichiers de la DSN évoluent d'un millésime à l'autre. Le statisticien peut modifier les règles utilisées dans le *mapping* pour gérer ces changements, mais indépendamment du modèle statistique qui peut rester inchangé. Pour les applications en aval, cette constance permet de comparer les données d'une année sur l'autre sans maintenance. Le modèle statistique lui-même peut évoluer, mais marginalement et de façon maîtrisée par le statisticien.

Cette étape de mapping requiert la construction préalable du modèle statistique appelé dans ARC « famille de norme ». Ce modèle permet de mettre en relation des entités statistiques et peut être soit défini directement dans l'application ou importé d'une spécification DDI. Le modèle de la DSN va contenir des tables, comme la table Employeur, avec les champs d'adresse de l'employeur, d'activité principale de l'entreprise ou le numéro d'immatriculation au répertoire Sirene[15] (Siret) ou encore la table Individu avec les champs du prénom ou du pays de résidence.

Tous ces modules de transformation peuvent s'appuyer sur des données externes intégrées par le statisticien, telles que des tables de nomenclatures, des référentiels ou des tables de correspondance. Par exemple, pour la DSN, les codes de pays de naissance des salariés sont recodés selon le code officiel géographique. ARC implémente ainsi en partie la phase de classification et codage.

15 Sirene : Système informatisé du répertoire national des entreprises et des établissements.

Mettre les fichiers de données à disposition pour les traitements ultérieurs

À l'issue du *mapping*, ARC finalise les fichiers de données. Les fichiers de la DSN, fraîchement transformés en tables de données exploitables, sont **mis à disposition** pour être **récupérés par les applications clientes**. L'application permet de gérer les clients de chaque source de données. Chaque livraison de données est horodatée et une seule récupération des données est autorisée par client, ceci afin d'éviter de doublonner des données. Une fois les fichiers téléchargés par les clients déclarés, ARC les supprime après un laps de temps défini par le statisticien.

Les données mises à disposition subissent d'autres transformations par les applications clientes de ARC : restructuration des données par unité statistique, calcul de variables dérivées, etc.

L'intégration des données administratives (Cotton et Haag, 2023) de la DSN passe par toutes ces étapes, de l'accueil réalisé par ARC jusqu'à leur transformation par les applications clientes. Le Répertoire Statistique des Individus et des Logements (Résil[16]) utilise le même schéma d'intégration et ARC a été choisi pour mettre en œuvre la fonction d'accueil.

▶ L'accueil des DSN et la volonté de réutilisation dans le système d'information de l'Insee

L'application informatique ARC a été conçue dans le cadre de la construction d'un Système d'Information sur l'Emploi et les Revenus d'Activité (SIERA) alimenté par diverses sources de données administratives. Il coordonne plusieurs applications et prend en charge l'ensemble des traitements produisant des indicateurs statistiques structurels et conjoncturels sur l'emploi et les salaires. Plus précisément, ARC avait pour objectif, au sein de ce SI, d'accueillir dès 2015 les fichiers mensuels de la Déclaration Sociale Nominative envoyés par la Cnav[17]. Ce fut un challenge pour la maîtrise d'ouvrage et l'équipe de développement : le flux des données à traiter mensuellement était massif et concentré (à l'origine, 1,8 million de fichiers par mois, reçus entre le 18 et le 22 de chaque mois). Les dessins de fichiers n'étaient pas stabilisés et les délais de mise en œuvre courts !

Deux besoins fondamentaux : performance et adaptabilité…

En matière de fonctionnalités, le produit devait répondre à deux besoins a priori orthogonaux : être suffisamment performant pour traiter tous les mois l'ensemble des fichiers en une semaine, (la contrainte telle qu'elle s'exprimait au début) et pouvoir s'adapter rapidement à des changements de ces fichiers. Pour cela, il doit être optimisé en permanence, tant sur le plan de la pertinence que de la rapidité des traitements. Cela implique souplesse et réactivité dans la mise au point des changements liés au contenu ou au format des données source, tout en minimisant l'impact sur les performances.

16 Voir l'article d'Olivier Lefebvre sur Résil dans ce même numéro.
17 Les données DSN sont produites par le GIP MDS (Groupement d'intérêt public pour la modernisation des déclarations sociales) *https://www.net-entreprises.fr/*.

> ***ARC doit être optimisé en permanence, tant sur le plan de la pertinence que de la rapidité des traitements.***

L'option prise est de laisser la main aux statisticiens pour programmer, tester et reprogrammer, dans un « bac à sable » les traitements d'accueil en fonction de l'évolution des sources et des attentes des traitements statistiques en aval. Par ailleurs, l'application est conçue de manière à ce que les réglages mis au point par les statisticiens ne s'appuient que sur des paramétrages des différentes phases de l'accueil, sans influer sur les traitements, et donc sans risque d'altérer les performances du système. Le statisticien peut ainsi se concentrer sur les aspects « métier » et son collègue en charge de l'exploitation informatique fait l'économie d'une phase d'optimisation.

… qui rendent l'application plus pérenne ?

Développer rapidement cette fonction d'accueil a permis la conception du projet ARC en mode agile[18]. D'un point de vue métier, ce besoin de souplesse était également très fort. En effet, les maintenances adaptatives pour absorber les changements de normes des fichiers source réalisées sur d'autres chaînes de traitement du SIERA comme le traitement de la N4DS[19], s'avéraient très coûteuses. ARC devait répondre à ce problème de façon générique pour être réutilisable et pouvoir accueillir d'autres sources que la DSN.

ARC fut ainsi déployé en 2015 pour l'accueil des fichiers XML de la DSN puis réutilisé dans le SIERA pour accueillir les fichiers des Déclarations Annuelles de Données Sociales (DADS), qui devaient « coexister » jusqu'en 2021 avec la DSN. L'application a également été utilisée pour l'accueil de fichiers déjà produits par l'Insee, dans une optique de mise dans un format commun. Dans cette première version, le produit était déjà capable de prendre en charge l'accueil de fichiers XML, CSV[20] et clé-valeur[21].

Les contraintes auxquelles il a fallu faire face pour développer un outil d'accueil de la DSN ont conduit à lui conférer les « bonnes propriétés » pour le désigner comme l'outil central d'un processus d'accueil des sources administratives découplé des traitements en aval. L'usage d'ARC s'est alors progressivement développé.

▶ Premières utilisations hors du cadre initial

Face à un changement de norme pour les liasses fiscales utilisées dans le processus Esane (Élaboration de statistiques annuelles sur les entreprises), il a été envisagé d'utiliser ARC plutôt que de développer un nouveau système spécifique à ce processus.

18 L'agilité a pour objectif d'orienter les efforts vers ce qui a le plus de valeur pour l'utilisateur, en s'adaptant aux changements à moindre coût.
19 N4DS : Norme pour les Déclarations Dématérialisées Des Données Sociales, utilisée par les DADS.
20 CSV désigne un format de fichiers dont le rôle est de présenter des données séparées par des virgules. Il s'agit d'une manière simplifiée d'afficher des données afin de les rendre transmissibles d'un programme à un autre.
21 Le format de stockage clé-valeur fait correspondre des clés (par exemple des rubriques métiers) avec des valeurs.

L'instruction technique menée a mis plusieurs points en évidence : la nécessité d'une évolution fonctionnelle d'ARC, puis l'intérêt de cet outil pour les utilisateurs et enfin le fait que l'utilisation de l'application représentait l'option la moins coûteuse pour prendre en charge cette évolution.

La couverture fonctionnelle d'ARC n'était cependant pas complètement suffisante pour prendre en charge les fichiers hiérarchiques complexes que sont les fichiers fiscaux. La normalisation du processus d'accueil proposé par l'application a néanmoins permis de réaliser les évolutions rapidement et d'enrichir l'offre du produit. ARC a donc pour la première fois été réutilisé hors SIERA en 2019 pour l'accueil des fichiers fiscaux dans Esane.

En 2020, dans le cadre d'un groupe de travail du système statistique européen sur le partage d'outils statistiques, ses fonctionnalités ont une nouvelle fois été étendues et cela dans deux directions. La première permettait de déployer l'application sur des infrastructures conteneurisées[22], permettant notamment une scalabilité[23] accrue. La seconde rendait possible l'appel des étapes d'accueil de fichier en mode web-service. ARC devenait alors capable de traiter des invocations à la demande, unitaires, en complément des traitements de masse initialement développés. SIRENE4 a ainsi intégré l'application dans ce mode d'utilisation de machine à machine pour contrôler de façon automatique la conformité des liasses d'immatriculation provenant du Guichet Unique[24] (Alviset, 2020).

ARC a donc évolué progressivement, pour devenir une application robuste et performante (***encadré***).

▶ Encadré. Caractéristiques techniques et performances.

ARC est un ETL opensource (*Extract Transform Load*) développé par l'Insee. Le code source de l'application est hébergé sur l'espace github inseeFr : *https://github.com/InseeFr/ARC*

ARC propose un module web, un module batch et un module web-service. Chaque module est conteneurisé avec Docker* et déployable de façon autonome selon les besoins métiers. Les conteneurs sont disponibles sur *dockerhub* : *https://hub.docker.com/u/inseefr*

Le module web propose une interface homme machine pour paramétrer, lancer des traitements sur des fichiers et piloter éventuellement les traitements massifs réalisés en batch. Le module batch permet de traiter des flux massifs de fichiers et d'assurer la reprise en cas d'erreur. Le module web-service expose un service de récupération de données et un service de traitement de fichier unitaire.

ARC peut traiter des fichiers XML, clé-valeur et de type texte, aux formats CSV, délimités ou positionnels.

Les performances d'ARC ont considérablement progressé depuis l'origine : les premiers chargements mensuels de la DSN réalisés en 2015 duraient 9 jours contre 60 heures aujourd'hui alors même que le volume de données à traiter est deux fois plus important. Cette amélioration significative a notamment été permise par le découplage entre stockage des données et traitements.

Les traitements dans ARC sont réalisés sur des bases PostgreSQL. L'instance dédiée au chargement de la DSN dispose actuellement d'une seule base de données avec 32 CPU et 32 Go de RAM.

La dernière version d'ARC est scalable horizontalement. Plutôt que de disposer d'une unique base de données avec beaucoup de ressources, l'application peut utiliser plusieurs petites bases de données en parallèle. Cette architecture permet d'éviter les problèmes de capacité de traitement inhérents à l'utilisation d'une seule machine : le temps de traitement décroît proportionnellement avec le nombre de bases de données dédiées à l'application.

* *Docker est un système permettant de créer, partager et exécuter des conteneurs.*

22 Une infrastructure de conteneurs permet d'automatiser le déploiement, la mise à l'échelle et la gestion des conteneurs. Un conteneur est un environnement d'exécution contenant tous les composants nécessaires (code, dépendances et bibliothèques) pour exécuter le code de l'application sans utiliser les dépendances de la machine hôte.
23 Capacité à s'adapter à des évolutions importantes du volume de données à traiter.
24 Le Guichet électronique des formalités d'entreprises (Guichet unique) est un portail internet sécurisé, auprès duquel toute entreprise est tenue de déclarer sa création ainsi que différents événements de vie, depuis le 1er janvier 2023.

Le changement d'échelle : l'organisation de la mutualisation

L'extension progressive des fonctionnalités et des usages d'ARC ainsi que son caractère incontournable dans différentes productions ont conduit à une prise en charge adaptée à ces enjeux. Il s'agissait d'une part de piloter la mise en œuvre et le déploiement d'investissements transverses (amélioration des performances, traitement des métadonnées, maintien en conditions opérationnelles et de sécurité, etc.), et d'autre part d'animer la communauté des utilisateurs (communication sur le produit et ses évolutions, recueil et priorisation des besoins, formations, etc.). Ces actions sont essentielles pour un produit mutualisé comme ARC afin de conserver un champ fonctionnel et de prendre en compte les besoins des divers utilisateurs.

En 2021, le programme Résil est devenu maîtrise d'ouvrage du produit ARC du fait de son positionnement central dans le système d'information, mais aussi en raison de la diversité des sources qu'il accueille.

Conclusion

Le monde de la donnée, comme celui de l'informatique, est en perpétuelle évolution. Pour répondre aux défis informatiques, les entreprises ont développé des stratégies comme la mise en place de l'agilité et du DevOps[25]. Cette idée a été reprise par l'univers de la data avec le développement du DataOps[26], qui vise notamment à concilier automatisation, reproductibilité, interactivité et traçabilité en matière de traitement des données, tout en réunissant différents métiers de la data autour d'outils communs. L'Insee, en proposant dans ses développements l'agilité, intègre déjà beaucoup de recettes du DataOps. Le service d'accueil que rend ARC s'inscrit pleinement dans cette démarche, de par la souplesse et le découplage des traitements qu'il met en œuvre.

ARC est une application qui a su évoluer au cours du temps en fonction de besoins multiples et variés. Une application désormais facile à déployer, adaptée à un chargement souple et performant de données externes, en amont d'applications statistiques. Une application qui devient centrale dans la mise en œuvre de la stratégie d'utilisation des données administratives visant à la fois la diversification des sources, la rapidité de leur traitement, la capacité à les mettre au service de processus statistiques différents, et cela en toute sécurité.

L'ouvrir vers des utilisations plus exploratoires, réalisés en « self service » par des statisticiens, peut permettre d'explorer plus facilement le potentiel de nouvelles sources de données, au service de l'innovation statistique.

25 Le DevOps est un mouvement en ingénierie informatique et une pratique technique visant à l'unification du développement logiciel et de l'administration des infrastructures informatiques.
26 Le DataOps est une méthode automatisée pour améliorer la qualité et réduire le temps de cycle de l'analyse des données. *https://dataopsmanifesto.org/fr/.*

▶ Bibliographie

- ALVISET, Christophe, 2020. La troisième refonte du répertoire Sirene : trop ambitieuse ou pas assez. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee. N°N4, pp 101-121. [Consulté le 16 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497083?sommaire=4497095*.

- BISEUL, Xavier, 2023. Data lineage : définition, principes et outils. In : *Journal du Net*. [en ligne]. 28 février 2023. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.journaldunet.fr/web-tech/guide-du-big-data/1516833-data-lineage-definition-principes-et-outils/*.

- BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 46-57. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168396?sommaire=4168411*.

- COTTON, Franck et DUBOIS, Thomas, 2019. Pogues, un outil de conception de questionnaires. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 17-28. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254216?sommaire=4254170*.

- COTTON, Franck et HAAG, Olivier, 2023. L'intégration des données administratives dans un processus statistique - Industrialiser une phase essentielle. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 104-125. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635829?sommaire=7635842*.

- DUBRULLE, Bertrand, ROSEC, Olivier et SUREAU, Christian, 2023. Une norme d'échange pour alimenter des référentiels et en assurer la qualité. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N°N9, pp 126-146. [Consulté le 18 juin 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635835?sommaire=7635842*.

- DONDON, Alexis et LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N°N9, pp 86-103. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.

- ERIKSON, Johan, 2020. Le modèle de processus statistique en Suède – Mise en œuvre, expériences et enseignements. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee. N° N4, pp. 122-141. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497085?sommaire=4497095*.

- 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*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 25-34. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3647025?sommaire=3647035*.

- LAMARCHE, Pierre et LOLLIVIER Stéfan 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N°N6, pp 28-46. [Consulté le 31 mai 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- RENNE, Catherine, 2018. Bien comprendre la déclaration sociale nominative pour mieux mesurer. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 35-44. [Consulté le 14 mars 2024]. Disponible à l'adresse : *https://www.insee.fr/fr/information/3647029?sommaire=3647035*.