Maintenance et Développement d'Application

Note de synthèse de l'étude préalable

Programme d'évolution du Siera
Projets de réingénierie
Projet Pirénés (ex-projet Accueil et traitement de la DSN) - Lot 1

Juin 2015

Équipe statistique : Charles Pilarski, Adeline Baudrey, Ourida Cherchem, Julie Roy, Jacques Traguany
Équipe informatique ; Manuel Soulier, Pierre Grimal, Matthieu Forestier-Blondeau, Sébastien Larat
Responsable de la démarche Maiol : Albane Gourdol

CREATION ET MISES A JOUR

Version	Date	Auteur	Pages modifiées	Objet de la mise à jour
0	Avril-mai 2015	Pierre Grimal Manuel Soulier Charles Pilarski		Initialisation des fiches
0.1	15/05/2015	Mylène Chaleix	toutes	Mise au format
0.2	18/05/2015	Pierre Grimal Manuel Soulier Charles Pilarski Sylvie Eghbal	toutes	Première relecture
0.3	19/05/2015	Pierre Grimal Manuel Soulier Charles Pilarski Sylvie Eghbal Mylène Chaleix	toutes	Version transmise au comité de pilotage Relecture division EFA
0.4	20/05/2015	Pierre Grimal		Compléments architecture applicative
1.0	01/06/2015	Charles Pilarski Sylvie Eghbal Mylène Chaleix	toutes	Version transmise N. Roth et la division Camap Relecture ensemble de équipes projet
2.0	12/06/2015	Charles Pilarski Mylène Chaleix	toutes	Version transmise au comité des investissements

STATUT

n° de version	date	approbation
0.3	19/05/2015	Envoi pour avis au comité de pilotage
1.0	01/06/2015	Version transmise à N. Roth et la division Camap
2.0	12/06/2015	Version transmise au comité des investissements

Sommaire

Avertissement

Le programme d'évolution du Siera repose sur trois investissements :

- la réingénierie des processus du Siera ;
- l'extension du champ de la publication et la mise en place d'une coproduction avec l'Acoss et la Dares pour les estimations trimestrielles d'emploi salarié ;
- la réorganisation du centre Statistiques sociales et locales (CSSL) du centre statistique de Metz en lien avec le transfert des travaux des pôles et des équipes de gestionnaires des directions régionales.

L'expression des besoins des projets de réingénierie[1] a été diffusée le 2 février 2015.

Compte-tenu des délais, il a été priorisé[2] la mise en place d'un premier projet pour être en mesure d'accueillir et traiter la nouvelle déclaration sociale nominative (DSN) dès son arrivée en février 2016.

Les objectifs de ce premier projet sont de réaliser les éléments indispensables permettant :

- d'assurer l'intégration de la DSN dans le système d'information sur l'emploi et de revenus d'activité ;
- de permettre aux gestionnaires du CSSL de traiter les déclarations DSN dès le premier semestre 2016 ;
- de créer au second semestre 2017 l'équivalent des fichiers « Tous salariés» sur la validité 2016.

Il a été retenu de découper le projet en deux lots qui correspondront à ces deux échéances :

- le lot 1, constitué du module d'accueil-réception mise au format, du module de traitements élémentaires et des sous-modules « interface avec les répertoires » et « gestion des nomenclatures » du module référentiel ;
- le lot 2, constitué d'une version minimale du module de traitements structurels, et de la composante « gestion des dégroupements » du module référentiel. En particulier le lot 2 reprend l'ensemble des fonctionnalités permettant de produire les informations nécessaires à la continuité de la production du produit « Tous salariés ».

À compter de cette étude préalable, le projet Accueil et traitement de la DSN est rebaptisé PIRÉNÉS (Projet Informatique de REfoNte sur l'Emploi et les Salaires). La présente note de synthèse d'étude préalable porte sur le premier lot du projet.

[1] Référence de l'expression des besoins transmise au comité des investissements : n°190/DG75-F201 du 1er février 2015

[2] Cf. § 3.2 de l'expression des besoins.

1 COMPLÉMENTS, ÉVOLUTION DU CONTEXTE, SOLUTIONS

1.1 Complément à l'expression des besoins

1.1.1 Rappel du positionnement du projet dans l'ensemble du système cible

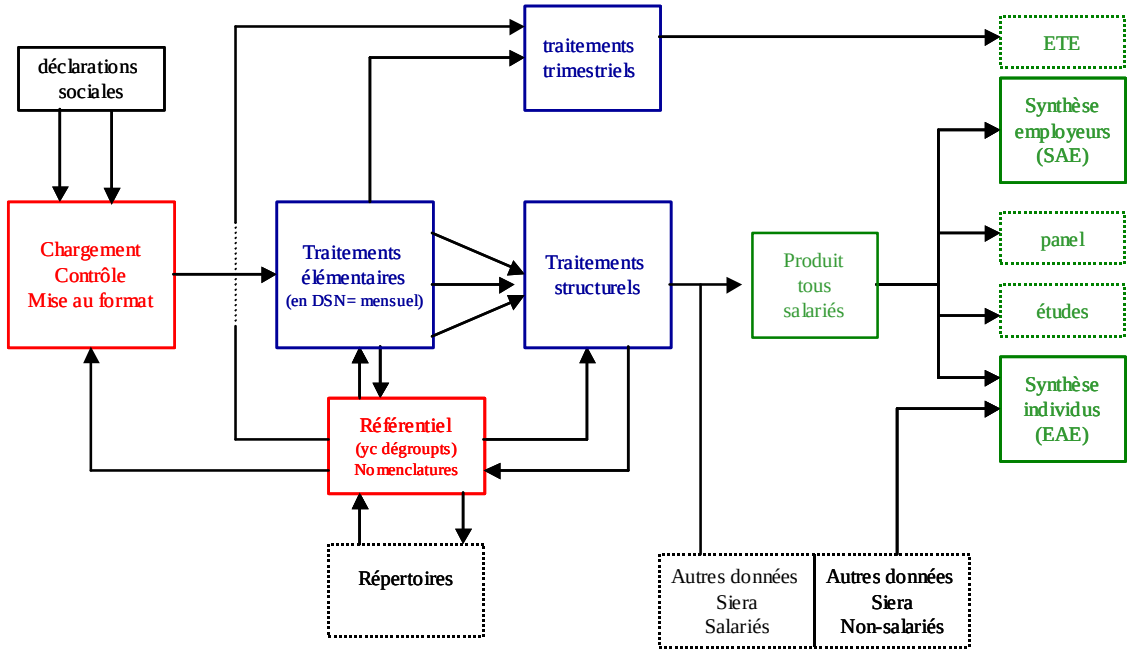
[extraits de l'expression des besoins]

Le système cible décrit dans l'expression des besoins reposait sur un découpage en trois blocs, eux-mêmes décomposés en plusieurs modules fonctionnels :

- un bloc d'infrastructures mutualisées pour le futur dispositif (module d'accueil-réception-mise au format, module Référentiel et gestion des nomenclatures) ;
- un bloc de traitements statistiques (module de traitements élémentaires ; module de traitements structurels, module de traitements trimestriels) ;
- un bloc de constitution des produits de diffusion (module de constitution du produit « tous salariés », module « synthèse sur les employeurs », module « synthèse sur les individus).

Le périmètre de ce système cible est inchangé par rapport à la description de l'expression des besoins.

Système cible Siera 2



Sa mise en place sera progressive et fera l'objet d'un découpage entre plusieurs projets. Chaque projet prendra en charge le développement d'un ou plusieurs modules fonctionnels, étant entendu que chaque module pourra faire l'objet de plusieurs versions successives en fonction des conclusions des travaux d'expertise statistique qui seront menés sur les sources alimentant le dispositif, en particulier sur la DSN. Le périmètre du lot 1 du premier projet, développé dans ce document, reste inchangé. Il mettra en place la première version des modules accueil-réception mise au format, traitements élémentaires, référentiel et gestion des nomenclatures. Ces modules pourront faire l'objet de nouvelles versions dans des projets ultérieurs.

Le périmètre du lot 1, celui du lot 2 et des projets suivants est résumé dans le tableau ci-après.

Périmètre fonctionnel du projet Pirénés et de l'ensemble des projets de réingénierie

		Projet Pirénés lot 1	Projet Pirénés lot 2	Projets ultérieurs
Bloc services transversaux	Accueil réception mise au format	x		
	Référentiel et gestion des nomenclatures	référentiel Sirus gestion nomenclatures	dégroupement v1 alimentation de Sirus	dégroupement v2
Bloc traitements statistiques	Traitements élémentaires	v1 sur déclaration DSN	v2, yc déclaration DADS amélioration identification et codage	v3, adaptation au champ de la fonction publique et des déclarations simplifiées
	Traitements structurels		v1 : traitements minimaux sur DSN et DADS	v2 : amélioration des contrôles et traitements de cohérence
	Traitements trimestriels (champ Insee)			x
Bloc constitution des produits de diffusion	Produit "tous salariés"		v1 (peut-être réutilisation des batch DADS)	v2 (traitement homogène N/N-1, amélioration des contrôles)
	Synthèse individus			x
	Synthèse employeurs			x
	Services (tirage d'échantillon, mise à dispo des données collectées, etc.)			x

1.1.2 Actualités de la DSN

Suivi du projet externe DSN

Le projet DSN , externe à l'Insee, est piloté par le groupement d'intérêt public « modernisation des données sociales » (GIP-MDS). Le dépôt des déclarations se fera via les portails *net.entreprises* (géré par le GIP-MDS) ou *msa.fr* pour le régime agricole. Le stockage des données sera pris en charge par deux opérateurs (Cnav et Acoss). C'est l'opérateur Cnav qui transmettra les données individuelles issues de la DSN à l'Insee.

Le décret d'obligation[3] pour les entreprises payant plus de 2 millions d'euros de cotisations[4] a porté ses fruits : 70% des entreprises concernées ont déposé une DSN en mai 2015, sachant que trois quarts le font au format de la phase 1 et un quart au format phase 2.

Phase 1[5] : la montée en charge se poursuit. Pour l'échéance du 5 mai 2015, environ 3,7 millions de salariés avaient fait l'objet d'une DSN mensuelle.

Phase 2[6] :

Bilan du pilote (prévu de novembre 2014 à février 2015, prolongé d'un mois) :

- une trentaine d'entreprises, sur 70 prévues dans le pilote, ont déposé une DSN avec succès ;
- aucune « grosse » entreprise n'a participé au pilote ;
- une seule entreprise de travail temporaire (ETT) est parvenue à déclarer mais de qualité insuffisante ;
- des alertes sur le RCD[7] (référentiel des cotisants et déclarants) avec des rejets à tort d'établissement.

La qualité des données est inégale :

- des problèmes de compréhension sur l'utilisation des blocs « changement » (Dares, Pôle emploi) ;
- 50 % de déclarations de bonne voire, d'excellente qualité ; 50 % de qualité médiocre, voire mauvaise en phase 2 (du point de vue du recouvrement, Acoss).

La mise en production de la phase 2 à compter du 17 mars 2015 a été prononcée par le comité directeur de la DSN du 6 mars 2015, mais sur un champ plus restreint (les travaux pilotes se poursuivent pour les ETT).

Prochaines échéances

Phase 2 : obligation de passage à la phase 2 en septembre 2015.

Phase 3 : pas de modification du calendrier :

- octobre 2015 - janvier 2016 : mise en place du pilote de la phase 3 ;
- février 2016 : mise en production de la phase 3.

Travaux entre l'Insee et le projet DSN

Janvier - février 2015 : **expression des besoins de l'Insee** et cahier des charges associé pour la phase 3 (variables demandées, contenu des flux, « tuyaux » entre la Cnav et l'Insee).

Points de vigilance pour l'Insee (vus lors de la réunion du 13 mai 2015 avec la maîtrise d'ouvrage stratégique du projet DSN) :

- travaux avec la direction de la sécurité sociale (décret phase 3 et arrêté filtre) et dossier Cnil pour la phase 3 : à lancer début juin 2015 ;
- veiller à être raccordé au pilote en octobre 2015.

[3] Décret n°2014-1082 du 24 septembre 2014.
[4] Soient environ 13 000 entreprises et 8 millions de salariés concernés.
[5] Périmètre de la phase 1 : substitution des déclarations de mouvement de main d'œuvre - DMMO, les déclarations de salaires pour les indemnités journalières - DSIJ, les attestations d'emploi - AE.
[6] Périmètre de la phase 2 : remplacement du BRC Acoss et intégration des entreprises de travail temporaire (ETT) avec le remplacement des relevés mensuels de mission (RMM) d'intérim.
[7] Le RCD est construit à partir du référentiel des employeurs et des indépendants de l'Acoss pour le régime général et de celui de la MSA pour le régime agricole.

1.1.3 L'organisation du projet

Les structures du programme et du projet sont en place (cf. annexe 2 pour la liste des réunions et les références des comptes-rendus).

1.2 Présentation générale des différentes solutions retenues

1.2.1 La solution retenue

Le lot 1 est constitué du développement des modules fonctionnels suivants :

- **accueil-réception mise au format**. Ce module fonctionnel a vocation à réceptionner les données administratives transmises par les partenaires, effectuer des contrôles de forme basique pour vérifier que les fichiers reçus sont bien conformes à l'attendu, et à transformer les données administratives en données statistiques élémentaires ;
- **traitements élémentaires**. L'objectif de ce module fonctionnel est de regrouper les travaux qui permettent de traiter les données statistiques élémentaires, sans qu'il y ait besoin de les croiser avec des données sur des périodes antérieures. Ce module permettra l'identification des unités statistiques d'intérêt et le codage des variables pouvant nécessiter l'intervention d'un gestionnaire ;
- **référentiel et gestion des nomenclatures**, pour ses sous-modules « référentiel et interface avec les répertoires » et « gestion des nomenclatures ». Ces deux composants seront réalisés dès le lot 1, le reste du module fonctionnel sera développé dans le lot 2 ou dans un projet spécifique en fonction de l'importance des développements à réaliser et des études statistiques préalables à conduire.

Le projet mettra en place deux applications :

- l'application ARC (Accueil Réception Contrôles), qui contiendra les fonctionnalités du module fonctionnel « accueil-réception mise au format », et celles du sous-module de « gestion des nomenclatures » ;
- l'application Artemis (Application de Reprise et de Traitements Élémentaires de l'eMploI et des Salaires), qui contiendra les fonctionnalités du module fonctionnel de traitement des données élémentaires, ainsi que celles du sous-module « référentiel et interface avec les répertoires ».

1.2.2 Les autres solutions étudiées

Une alternative « métier » : un nouveau frontal

Une alternative à la refonte totale du Siera a été décrite dans l'expression des besoins : recréer à partir des DSN mensuelles un équivalent des DADS administratives et des BRC, ce scénario permettant de maintenir en l'état les chaînes de traitement actuelles.

Cette alternative a été écartée car :

- elle ne permet pas de remplir les objectifs stratégiques de la maîtrise d'ouvrage[8], en particulier ceux de mieux lisser la charge de travail des équipes de gestionnaires et d'améliorer les délais de production ;
- la comparaison entre les données annuelles et les données conjoncturelles devra être redéfinie (puisqu'il n'y aurait plus de toutes façons qu'une seule source de données en entrée des processus) ;
- cette solution est peu pérenne compte tenu des évolutions récurrentes portant sur les déclarations sociales, et des contraintes techniques sur les applications existantes qui ne peuvent que difficilement évoluer ;
- elle peut conduire à créer des problèmes de qualité dans les données liées au recodage, dont l'analyse est coûteuse pour les équipes métier en production courante.

[8] Cf. § 2.2 de l'expression des besoins.

Il n'a pas été étudié dans le cadre du lot 1 d'autre solution métier alternative.

Alternative technique sur le stockage des données : relationnel versus *Big Data*

Les premières volumétries annoncées en début d'étude préalable, autour de 30 To, ont amené l'équipe projet à envisager l'utilisation d'une base de données (BDD) de type *Big Data*. Le *Big Data* est en effet capable de gérer de très gros volumes de données ainsi que de grandes variétés de données. Les vitesses d'accès et de tabulation/restitution sont également très élevées.

Le *Big Data* à l'Insee

Actuellement à l'Insee, le *Big Data* est une nouveauté au stade, maintenant bien avancé, d'expérimentation dans le cadre du projet de "Données de Caisses". Il n'y a pas ou peu d'équipes formées au *Big Data* aussi bien en développement qu'en suivi d'exploitation.

Le CEI serait en mesure de mettre en place des plateformes avec des BDD *Big Data* pour l'automne 2015. Ce calendrier n'est pas compatible avec celui du projet Pirénés, lequel demande que les premières mises en production interviennent fin 2015.

Un choix raisonnable : la BDD relationnelle

L'application ARC et le poste de pilotage associé devront être opérationnels en production dès la fin 2015 (première campagne réelle en février 2016). Les équipes actuelles ont de fortes compétences en BDD relationnelles. Le passage d'Oracle à PostgreSQL ne remet pas fondamentalement en cause les acquis des personnes.

La volumétrie des données et l'organisation en bases de données ont été affinées tout au long de l'étude préalable, ce qui a permis d'envisager le recours à des BDD relationnelles. En effet, l'estimation initiale de la volumétrie postulait que l'ensemble des données serait stocké en base de données alors que seule une petite partie (1/60) aurait vraiment été sollicitée par les processus métier. En repensant les processus de conservation des données sous forme de fichiers compressés dans des espaces, la dernière estimation des volumes de bases de données a été ramenée à deux bases de 3,5 To ce qui justifie d'autant plus le choix d'une BDD relationnelle.

En conclusion, la volumétrie cible, les délais imposés par le programme et le degré de maturité du *Big Data* à l'Insee ont rapidement orienté le choix vers une BDD relationnelle PostgreSQL.

2 DESCRIPTION DE LA SOLUTION RETENUE

2.1 Description métier

2.1.1 Description générale du processus métier

Principaux objectifs des modules fonctionnels développés dans le lot 1

Le Siera 2 a été découpé en plusieurs modules fonctionnels, de façon à rendre compte des besoins métiers et à partager au mieux les développements à mener et les projets à lancer.

Le lot 1 du projet Pirénés prend en charge le développement des modules fonctionnels suivants :

- le module d'accueil-réception mise au format. Ce module fait partie des modules d'infrastructure du Siera 2, dans le sens où si son premier objectif est bien de fournir une solution d'accueil pour la déclaration sociale nominative(DSN), il a vocation à accueillir tout ou partie des déclarations sociales en entrée du Siera 2 lorsque le développement de celui-ci sera achevé. La version de ce module fonctionnel mise en production à l'horizon de la fin du lot 1 sera quasi-définitive ;
- le module de traitements élémentaires. Ce module est spécifique au traitement des données issues de la déclaration sociale nominative (DSN). Toutefois il pourrait facilement être dupliqué ou adapté pour le traitement d'autres sources de données administratives. Le module de traitements élémentaires est l'un des trois modules fonctionnels de traitements statistiques du Siera 2. Il sera complété par un module de traitements structurels (lot 2 du projet), puis par un module de traitements trimestriels dans le cadre d'un projet ultérieur. L'objectif de ce module est de permettre le traitement au fil de l'eau des déclarations sociales nominatives, pour les variables sur lesquelles un tel traitement est plus efficace qu'un traitement annuel. À ce stade du projet, les traitements d'identification des salariés et des employeurs, ainsi que les codages des libellés de profession/catégories sociales (PCS) et des communes de lieu de résidence des salariés, apparaissent comme permettant de lisser les charges pour les équipes de gestionnaires. La version mise en production à l'issue du lot 1 de ce module fonctionnel ne correspondra pas à la version définitive, un certain nombre de traitements (en particulier sur le codage de la PCS) devant faire l'objet d'investigations complémentaires dès lors que la connaissance de la DSN se sera affinée à l'issue de la mise en production du lot 1 par les équipes métiers, appuyées en tant que de besoin par les équipes projet ;
- le module « référentiel et gestion des nomenclatures ». Ce module permet de gérer les liens et la cohérence entre le Siera et le répertoire statistique sur les entreprises et les établissements Sirus. Il s'agit d'un module d'infrastructure du Siera 2, puisqu'à terme il sera la référence de chaque application du Siera 2 pour le traitement de l'exhaustivité du champ, pour la récupération des données complémentaires sur les employeurs, et le dégroupement des unités employeuses. La majeure partie de ce module ne sera développée qu'au cours du lot 2 du projet Pirénés. La version mise en production dans le cadre du lot 1 ne contiendra que les fonctionnalités propres à la gestion des nomenclatures, à la gestion de base du référentiel (ajout, suppression, modification d'unités) et celles liées à l'interrogation de ce référentiel (test sur l'existence d'une unité, rapatriement des caractéristiques d'une unité pour une validité donnée).

Les travaux et le calendrier de production cible

Le processus-cible du lot 1 du projet Pirénés s'organise autour de deux types de travaux :

- les travaux relevant des infrastructures du Siera 2, dont le calendrier dépend des fournisseurs internes ou externes. C'est le cas du processus de chargement des données du module d'accueil-réception-mise au format, qui est mis en œuvre lorsque de nouvelles données sont transmises par les partenaires. C'est également le cas du processus de mise à jour du référentiel, qui est complété dès qu'un nouveau référentiel Sirus est disponible. Les modules d'infrastructure agissent

comme un sas dans lequel sont traitées puis retenues sous une forme utile à Siera 2 les données provenant de l'extérieur.

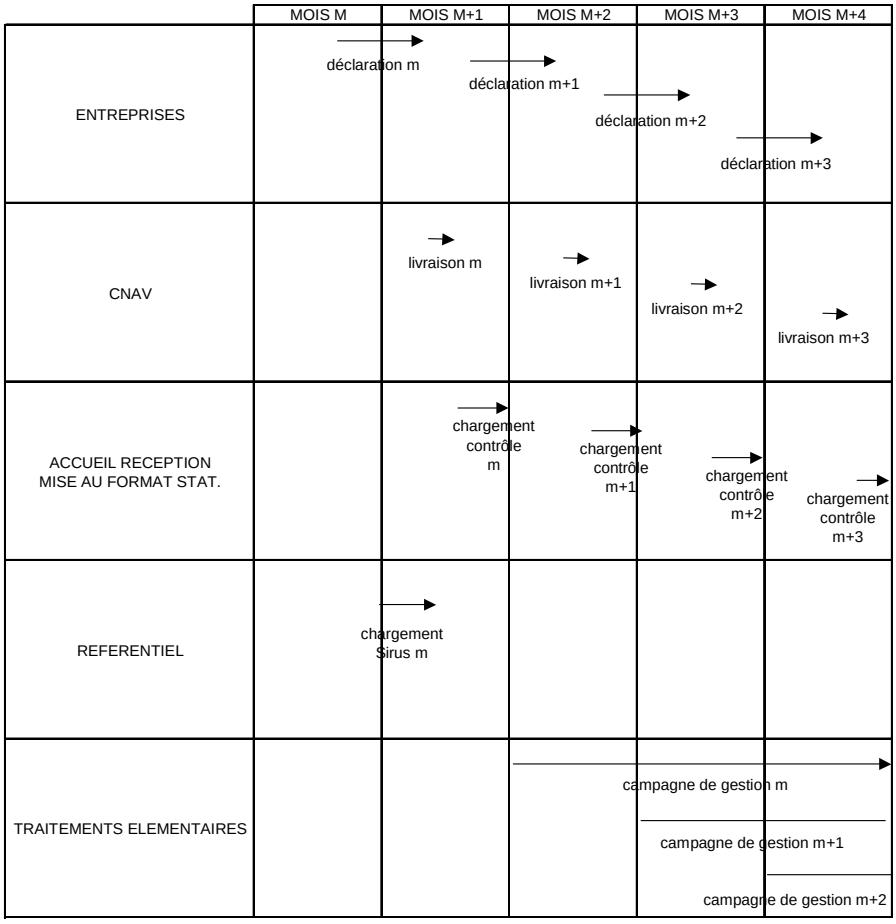
- les travaux relevant des applications métier du Siera 2, qui sont rythmés par la récupération de données en provenance des infrastructures[9] selon un calendrier de gestion défini à l'avance, et paramétrable par des acteurs métiers dans les applications. C'est le cas de chaque sous-processus du module de traitements élémentaires, ainsi que des sous-processus de mise à disposition des données qui sont présents dans chaque module d'infrastructure.

Le calendrier de production sera largement dépendant des dates de mise à disposition des données par les fournisseurs (Cnav dans le cas de la livraison des données de la DSN, Driss dans le cas de la livraison des référentiels standards de Sirus).

Les hypothèses retenues pour définir le calendrier de production ci-dessous sont les suivantes :

- livraison des données de la déclaration sociale nominative par la Cnav pour la validité M entre le 18 et le 25 du mois M+1 ;
- livraison d'un référentiel Sirus trois fois par an (au début des mois de février, juin et octobre) ;
- ouverture des campagnes de traitements mensuels (automatiques et de reprise manuelle par les gestionnaires) à J+15 après la livraison des données de la Cnav ;
- fermeture des campagnes de traitements mensuels 3 mois après la date d'ouverture (chaque campagne devrait se dérouler sur trois mois glissants, afin de gérer les périodes d'absence des gestionnaires et pour profiter de la fusion des cas de reprise similaires entre campagnes actives).

Calendrier de production de la DSN et de traitement par l'INSEE



2.1.2 Module fonctionnel Accueil-réception-mise au format

Objectifs du module

Le module d'accueil-réception-mise au format constitue la première étape du traitement des données de la future chaîne applicative Siera 2. L'objectif de ce module est de proposer une solution d'accueil des données administratives stable même lors des changements de norme de la déclaration administrative, d'effectuer un certain nombre de contrôles afin de vérifier la qualité des données, et d'assurer la transformation de ces données dans un schéma exploitable par la suite par les applications clientes de traitements statistiques.

Ce module doit permettre de charger et d'exploiter n'importe quel type de donnée administrative. Le paramétrage du chargement (fichier plat ou fichier XML), la reconnaissance des normes associées à des déclarations, les contrôles appliqués aux données et la transformation des données administratives en données statistiques (le « mapping ») peuvent être modifiés et améliorés au cours du temps par l'interactif, soit par le responsable informatique de l'application (RIA), soit par l'administrateur d'application (AA) lui-même, qui disposent alors d'un cadre de travail pour faire évoluer les paramètres sans toucher au cœur de l'application.

Compte tenu des possibilités de modification des traitements directement par l'interactif et des risques qu'une mauvaise spécification pourrait avoir sur la production et étant donné les volumes de données très importants à traiter, les acteurs qui l'utilisent doivent disposer d'espaces de tests similaires à l'espace de production mais distincts, afin de tester, qualifier les nouvelles règles spécifiées et mesurer leur impact sur les données chargées avant leur mise en production.

Les principaux objectifs du module fonctionnel sont les suivants :

- proposer une solution d'accueil et de chargement en base des données administratives, quel que soit leur format ou leur destination dans le futur Siera ;
- identifier la source des données chargées (norme et version de la déclaration, validité de la déclaration) et réaliser des contrôles afin de vérifier l'adéquation entre les données transmises et les variables et modalités attendues ;
- transformer les données administratives contrôlées en données statistiques, c'est-à-dire réaliser des transformations des données de la source vers un ensemble de variables statistiques stables ;
- permettre à l'administrateur d'application et au RIA de modifier, compléter ou supprimer les paramètres transmis à la chaîne de production, de la réception des fichiers jusqu'à leur mise à disposition des applications clientes ;
- alimenter les applications clientes de traitements statistiques avec les données statistiques élémentaires, sur le périmètre fonctionnel qui leur est dévolu ;
- fournir un environnement sécurisé permettant de tester les modifications apportées aux paramètres avant leur mise en production, dans un environnement équivalent à celui de production.
- permettre aux utilisateurs de l'application de consulter, modifier et charger/recharger des fichiers de données.

Les données administratives seront conservées sur une durée limitée : la durée de conservation des données administratives transmises par les partenaires sera d'un an à compter de leur réception, et de trois ans pour les données « normées » (ie après chargement en base et identification de la norme et de sa version). La conservation des données administratives transmises par les partenaires est nécessaire afin de corriger et recharger des fichiers qui s'avèreraient erronés. La conservation des données « normées » est nécessaire afin de pouvoir mettre ponctuellement à disposition des unités métier, les variables chargées mais non-exploitées par les applications clientes, afin de les étudier et le cas échéant, d'améliorer les traitements statistiques.

[9] modules fonctionnels 'Référentiel' et 'Accueil-réception-mise au format'.

Enfin en tant que module d'infrastructure, il permettra de gérer les nomenclatures et tables de passage métiers utiles aux traitements dans les applications clientes. Ces métadonnées seront transmises aux processus clients en même temps que les données elles-mêmes.

Flux d'information

En entrée, le module est alimenté par deux types de données :

- les données des déclarations transmises par les partenaires et déposées automatiquement sur des espaces de réception dédiés à un fournisseur et un type de donnée ;
- les données corrigées ou créées par les utilisateurs, déposées via l'interactif dans les espaces de réception pertinents.

Dans un premier temps, deux espaces de chargement seront mis en place : un espace pour les données DSN au format XML transmises par la Cnav ; un espace pour les données DADS issues du Frontal N4DS. Les deux espaces de réception de l'application pourront être complétés par l'équipe projet ou le futur RIA de l'application par d'autres espaces si des fichiers de format différent devaient être amenés à être utilisés.

En sortie, deux types de données seront générés :

- les données statistiques élémentaires, c'est-à-dire le résultat de la transformation des variables déclarées en variables statistiques (recodifications, filtrage, re-calculs *ad hoc*). Ces données élémentaires, dites « mappées », sont stockées dans des bases de données, et pourront être transférées aux processus clients via un webservice. Dès lors qu'elles auront été récupérées par le processus client, elles seront transférées dans les espaces de conservation (pour une durée d'un an) et supprimées des bases de données en production ;
- les données administratives chargées à l'état « normé », mises à disposition des équipes métier sous AUS en vue d'exploitations ponctuelles, pour une durée de trois ans.

Acteurs

En dehors des acteurs techniques qui mettent à disposition les données transmises par les partenaires (SEF), ou ceux qui récupèrent les données pour leurs traitements (applications clientes, dont le module fonctionnel « traitements élémentaires »), seuls l'administrateur d'application (sous la responsabilité de la division EFA) et le RIA auront accès à l'applicatif et ses fonctionnalités. Les deux acteurs disposent tous deux de toutes les fonctionnalités de l'interactif. Un mode de travail collaboratif devra être mis en place entre ces deux acteurs.

Traitements

Le processus cible du module fonctionnel est subdivisé en deux sous-processus.

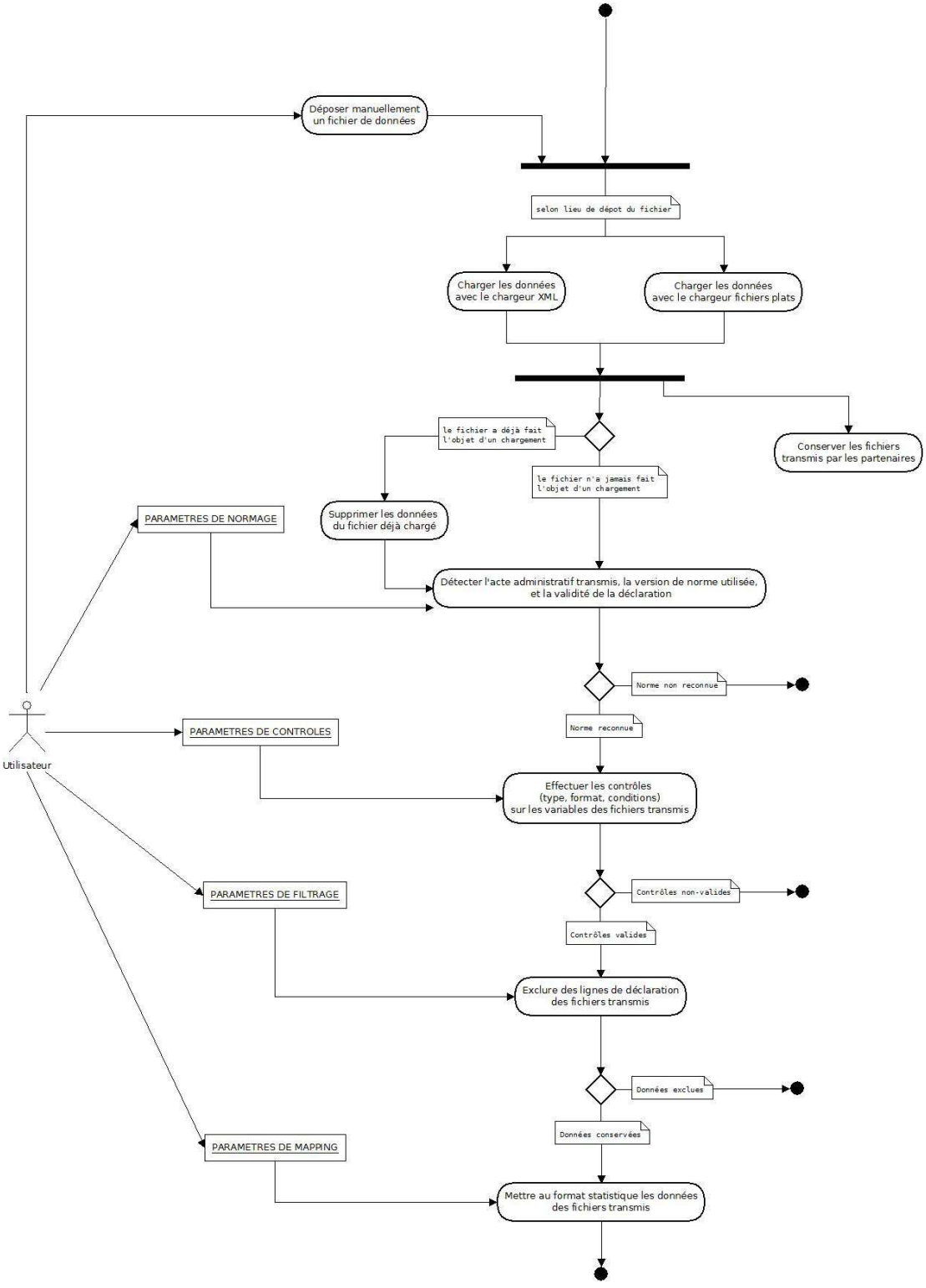
Le sous-processus « charger les données » prend en charge les traitements depuis la réception des données administratives jusqu'à leur mise au format statistique. Ce sous-processus se subdivise lui-même en quatre cas d'utilisation principaux :

- le chargement en base des données reçues à l'aide du chargeur adéquat. Ce chargeur est déterminé en fonction de l'espace où a été déposé initialement le fichier à charger ;
- la phase de « normage », qui consiste à identifier pour chaque fichier reçu l'acte administratif déclaré, la version de la norme d'échange utilisée et la validité de déclaration du fichier. Les fichiers dont la norme et/ou la validité n'ont pas pu être déterminés sont exclus de la suite des traitements et devront être repris par l'administrateur d'application ;
- la phase de contrôle de données. Les contrôles sont spécifiques à un type de déclaration et une version de norme. Ils peuvent faire l'objet de versions successives pour une même version de la norme selon les validités de déclarations. Il existe trois types de contrôles : les contrôles de cardinalité entre variables, les contrôles de format et les contrôles de condition sur la valeur des variables. Si une règle de contrôle n'est pas satisfaite, la ligne de déclaration est exclue mais pas le reste du fichier. Lorsqu'un fichier comporte un nombre significatif de lignes de déclarations en

anomalie, l'ensemble du fichier est exclu de la suite du traitement et devra faire l'objet d'une correction manuelle ;

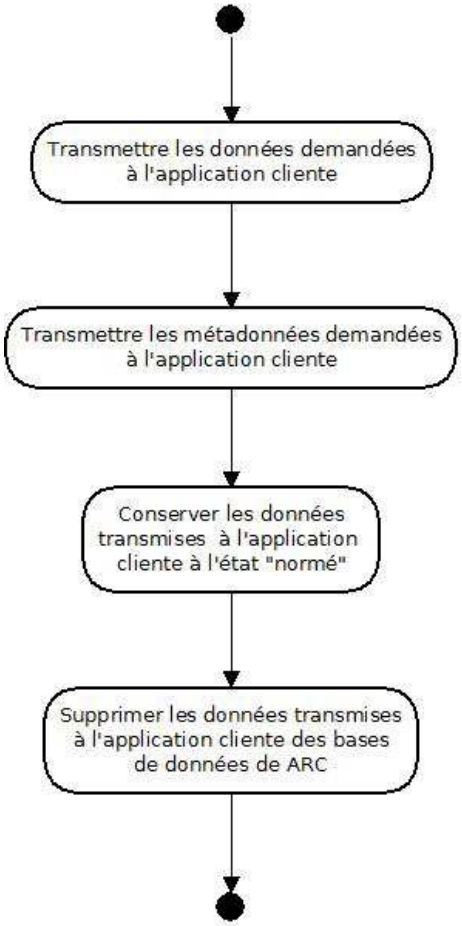
- la phase de filtrage ou d'exclusion. À cette étape, une partie des données peuvent être exclues si une partie des variables ou du champ déclaré ne doit pas être transmis aux applications clientes ;
- la phase de « mapping » ou de mise au format statistique. Les données non-exclues sont transférées, parfois après transformations, dans des tables métier qui correspondent aux objets qui seront transférées aux applications clientes.

Chacune de ces phases fait l'objet d'un paramétrage par l'utilisateur de l'application : la façon de détecter la version de la norme et la validité, les jeux de contrôles, les règles d'exclusion ou encore les règles de « mapping » sont autant de paramètres qui sont spécifiés via l'interactif par les utilisateurs.


Déposer manuellement un fichier de données
selon lieu de dépot du fichier
Charger les données avec le chargeur XML
Charger les données avec le chargeur fichiers plats
le fichier a déjà fait l'objet d'un chargement
le fichier n'a jamais fait l'objet d'un chargement
Conserver les fichiers transmis par les partenaires
PARAMETRES DE NORMAGE
Supprimer les données du fichier déjà chargé
Détecter l'acte administratif transmis, la version de norme utilisée, et la validité de la déclaration
Norme non reconnue
Norme reconnue
Utilisateur
PARAMETRES DE CONTROLES
Effectuer les contrôles (type, format, conditions) sur les variables des fichiers transmis
PARAMETRES DE FILTRAGE
Contrôles non-valides
Contrôles valides
Exclure des lignes de déclaration des fichiers transmis
Données exclues
Données conservées
PARAMETRES DE MAPPING
Mettre au format statistique les données des fichiers transmis

Le sous-processus « transmettre les données aux applications clientes » prend en charge :

- la transmission des données « mappées » sur le champ défini par les applications clientes ;
- la conservation des données normées et leur mise à disposition aux utilisateurs sur des espaces sécurisés ;
- le nettoyage des bases de données, c'est-à-dire la suppression des données issues des fichiers transférés de toutes les bases de données.



2.1.3 Module fonctionnel Traitements élémentaires (version 1)

Objectifs du module

Le module fonctionnel de traitements élémentaires se situe en aval du module d'accueil-réception-mise au format qui réceptionne, contrôle et met les données administratives dans un format statistique stable. Il se situe en amont du module de traitements structurels (développé au lot 2 du projet Pirénés), dont l'objectif sera de réaliser, à une périodicité *a priori* annuelle, les traitements de cohérence et le cas échéant de redressement des variables de salaire et de durée.

Les objectifs principaux de ce module sont les suivants :

- récupérer à des échéances définies par les équipes métier les données statistiques brutes issues de la DSN via le module d'accueil-réception-mise au format ;
- réaliser de façon automatique ou manuelle l'identification (employeur ou salarié[10]) et le codage des principales variables d'intérêt (PCS, lieu de résidence du salarié) ;
- consolider ces données avec l'ensemble des données passées déjà traitées afin d'obtenir une base de données homogène contenant l'ensemble des données ayant été traitées par le module.

Afin de limiter la charge de reprise manuelle, l'application devra permettre de moduler la charge de reprise en fonction des ressources disponibles dans les équipes de gestionnaires, et mettre en œuvre un dispositif d'apprentissage des traitements gestionnaires et/ou de l'automate.

Flux d'information

En entrée du processus,: à chaque initialisation de campagne, le module récupère d'une part les données de la DSN mises au format statistique à l'issue du module d'accueil-réception-mise au format, et d'autre part toutes les métadonnées (nomenclatures, tables de passage) nécessaires au codage.

En cours de processus, l'automate utilise un certain nombre de services externes à l'application afin d'identifier les principales unités statistiques (salarié, employeur) et le cas échéant de faire le codage de certaines données d'intérêt. Pour ce faire, le traitement mobilise les services suivants :

- le module référentiel employeur du Siera, pour l'identification des unités employeuses. Ce module est alimenté essentiellement par le répertoire statistique Sirus ;
- le service d'identification Siam-Sirene, pour l'identification des unités absentes du référentiel Siera. Lorsqu'un appel Siam donne lieu à une identification, cet employeur est également ajouté dans le référentiel Siera.
- le service d'identification à la BRPP, pour les salariés dont le NIR et/ou la date de naissance n'aurait pas été certifiés par le partenaire ;
- le service de codage « Sicore-commune », pour le codage automatique du lieu de résidence des salariés à partir des données déclarées (essentiellement code postal et libellé de la commune de résidence) ;
- le service de codage « Sicore-PCS », pour effectuer le codage automatique de la PCS à partir des variables déclarées (libellé de poste, caractéristiques de l'employeur) en vue de la comparer avec le code PCS déclaré par l'employeur et le cas échéant se substituer à ce dernier.

En sortie de processus, les données produites alimenteront une fois par an le module de traitements structurels.

[10] L'identification des salariés ne fera pas l'objet de reprise manuelle.

Acteurs

Le module de traitements élémentaires est doté de plusieurs profils utilisateurs. Les profils sont imbriqués les uns dans les autres, en partant de l'administrateur d'application qui dispose de l'ensemble des droits, jusqu'au gestionnaire qui dispose des droits les plus restreints.

Le gestionnaire a accès aux écrans de reprise : identification de l'employeur, reprise de la commune de résidence, reprise de la PCS. Il n'a accès qu'à une sous-partie de l'ensemble des cas à traiter, qui s'entend comme un portefeuille de cas de reprises qui lui sont attribués *ex-ante* par le chef d'équipe.

Le chef d'équipe dispose, en plus de l'accès aux écrans des gestionnaires, d'un écran lui permettant de gérer les portefeuilles de reprise des agents dont il est hiérarchiquement (au sens de l'application) responsable. Cet écran est complété par un poste de suivi des travaux de reprise des gestionnaires.

L'expert dispose, en plus des écrans accessibles au chef d'équipe, d'un écran lui permettant de valider, exporter et importer les règles de capitalisation de masse et/ou unitaires (cf. infra, paragraphe « Le processus de capitalisation »). Ainsi il peut ajouter des règles, en supprimer ou en modifier.

L'administrateur d'application a accès aux écrans de gestion des profils utilisateurs (ajout, suppression, modification des profils pour chaque agent), à l'écran de gestion des calendriers de campagne (ajout, modification, suppression d'une campagne), ainsi qu'à l'écran pré-campagne qui permet d'une part de visualiser et de filtrer les cas qui seront effectivement visibles en reprise, d'autre part d'ouvrir effectivement la reprise aux gestionnaires.

Les traitements et les interfaces

Le processus de production est rythmé par la notion de campagne. Une campagne se caractérise par une date d'ouverture, c'est-à-dire le moment où les données sont récupérées pour accueil, contrôle et mise au format statistique, une validité, une date de clôture de la campagne, c'est-à-dire la date à laquelle les données traitées automatiquement par l'application ou manuellement par les gestionnaires ne seront plus modifiables et seront consolidées avec les données des campagnes précédentes.

Il est possible d'avoir simultanément plusieurs campagnes actives[11], à condition que leurs validités se succèdent strictement. Les cas de reprises propres à chaque campagne sont alors fusionnés afin qu'un cas apparaissant sur plusieurs campagnes actives n'ait à être traité qu'une seule fois.

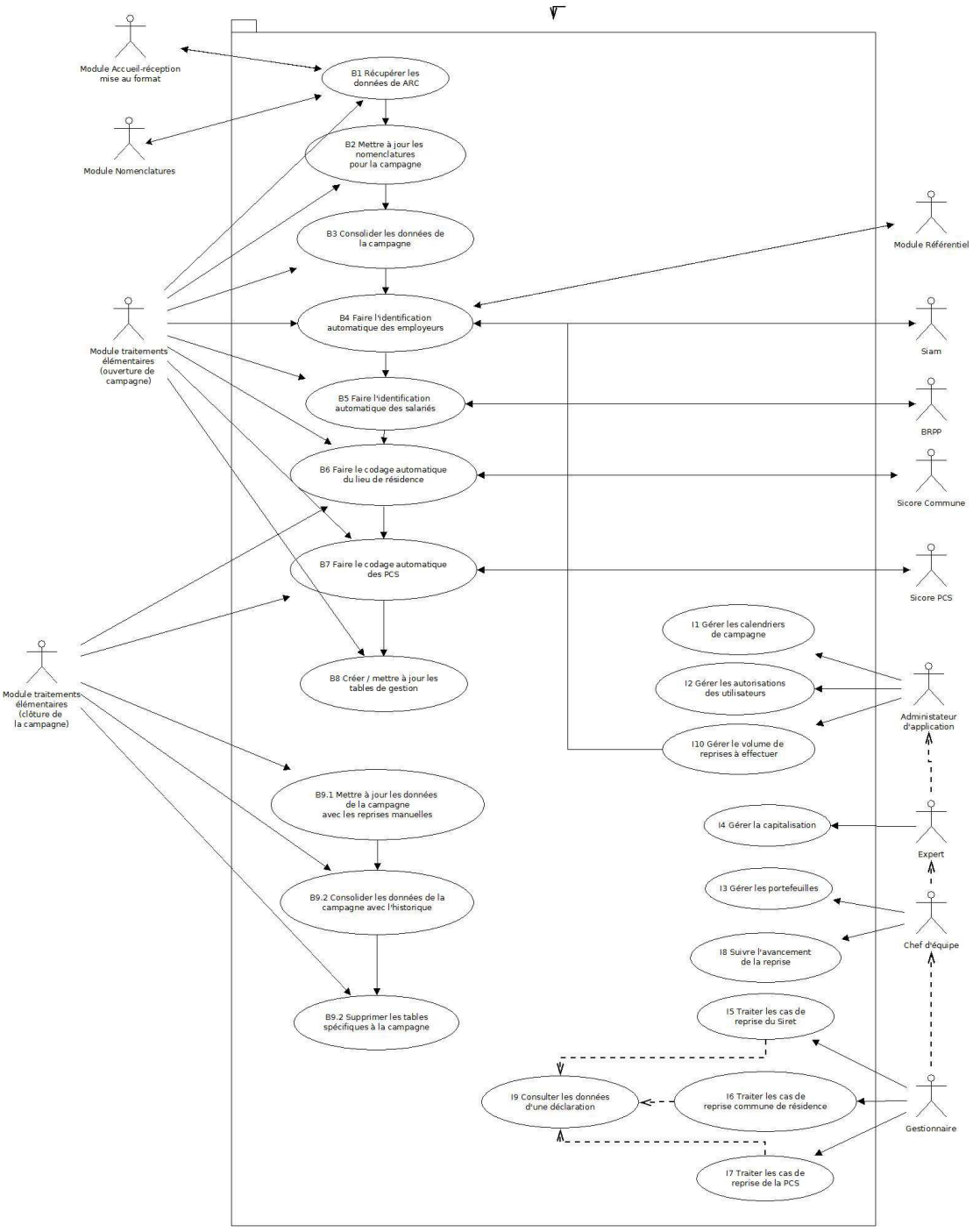
Les traitements de chaque campagne sont décomposés en trois phases :

- la phase d'initialisation de la campagne, dont l'objectif est de récupérer les données élémentaires mises au format statistique et les nomenclatures, de faire l'ensemble des identifications et des codages automatiques, puis d'identifier les cas à soumettre aux gestionnaires. À l'issue de cette phase, l'administrateur d'application décide d'ouvrir la campagne aux gestionnaires, c'est-à-dire de rendre visibles les cas à reprendre. Il peut également les filtrer sur des règles de priorisation définies par l'automate si le volume de reprises est trop élevé ;
- la phase de codage, au cours de laquelle les gestionnaires effectuent les reprises des cas identifiés à la phase précédente ;
- la phase de clôture de la campagne, qui recalcule les données à coder à partir des traitements et règles de codage les plus récents, consolide l'ensemble des données de la campagne (codées automatiquement ou traitées par les gestionnaires) avec les données des campagnes précédentes, et purge les tables créées uniquement pour la campagne.

[11] Une campagne est active si elle a déjà été ouverte et n'a pas encore été fermée.

Les interfaces de la future application permettront de répondre à trois types de besoins :

- les interfaces permettant la reprise par les gestionnaires des cas d'identification et de codage :
  - l'interface de reprise de l'employeur (ou d'identification du Siret) ;
  - l'interface de reprise de la PCS ;
  - l'interface de reprise de lieu de résidence des salariés ;
  - l'interface de consultation des déclarations à traiter dans la campagne ;
  - l'interface de consultation de l'historique des capitalisations.
- les interfaces de pilotage de la campagne de gestion :
  - l'interface de gestion des calendriers de campagne et de pilotage de la campagne ;
  - l'interface d'affectation des portefeuilles de reprise entre les différents acteurs ;
  - l'interface de suivi d'avancement de la campagne, qui présentera des indicateurs de taux de reprise par acteur pour la campagne.
- les interfaces d'administration d'application et d'expertise :
  - l'interface de gestion des droits des utilisateurs ;
  - l'interface de gestion des tables de capitalisation, qui permettra d'invalider des règles de capitalisation générées par l'automate ou par des gestionnaires, ou bien d'ajouter des règles nouvelles.


Module Accueil-réception mise au format
Module Nomenclatures
Module traitements élémentaires (ouverture de campagne)
Module traitements élémentaires (clôture de la campagne)
B1 Récupérer les données de ARC
B2 Mettre à jour les nomenclatures pour la campagne
B3 Consolider les données de la campagne
B4 Faire l'identification automatique des employeurs
B5 Faire l'identification automatique des salariés
B6 Faire le codage automatique du lieu de résidence
B7 Faire le codage automatique des PCS
B8 Créer / mettre à jour les tables de gestion
B9.1 Mettre à jour les données de la campagne avec les reprises manuelles
B9.2 Consolider les données de la campagne avec l'historique
B9.2 Supprimer les tables spécifiques à la campagne
Module Référentiel
Siam
BRPP
Sicore Commune
Sicore PCS
I1 Gérer les calendriers de campagne
I2 Gérer les autorisations des utilisateurs
I10 Gérer le volume de reprises à effectuer
Administateur d'application
I4 Gérer la capitalisation
Expert
I3 Gérer les portefeuilles
I8 Suivre l'avancement de la reprise
Chef d'équipe
I5 Traiter les cas de reprise du Siret
I9 Consulter les données d'une déclaration
I6 Traiter les cas de reprise commune de résidence
I7 Traiter les cas de reprise de la PCS
Gestionnaire

Le processus de capitalisation

Un dispositif de capitalisation au fil de l'eau est mis en place afin de limiter les cas de reprise transmis aux gestionnaires et le volume des appels aux services externes (Siam, BRPP, Sicore-Commune, Sicore-PCS).

La capitalisation repose sur la constitution d'un lien entre une ou plusieurs variables identifiantes et une variable identifiée. Par exemple, le codage du lieu de résidence d'un salarié dans une déclaration repose sur une clé mettant en relation, un salarié, un code postal déclaré et un libellé de commune déclaré d'une part, un code commune au sens du code officiel géographique d'autre part. La combinaison de l'ensemble des variables identifiantes détermine la variable identifiée de façon unique. Dans ce cas, on parlera de capitalisation unitaire, dans le sens où chacune des variables identifiantes doit être renseignée pour que le lien soit créé.

A contrario, il est possible d'établir des règles de reprise de masse, pour lesquelles une ou plusieurs variables identifiantes ne seront pas décrites. Par exemple, on pourra définir une règle sur le nom de commune uniquement, indépendamment du code postal, de l'employeur déclarant ou du salarié. Ainsi l'affectation du code commune s'appliquera à tous les enregistrements en échec de codage pour lesquels le libellé de commune aura été détecté. On parlera dans ce cas de capitalisation de masse, dans le sens ou plusieurs enregistrements de nature différente pourront être affectés par une unique règle.

Les règles de capitalisation de masse et celles de capitalisation unitaires peuvent cohabiter et parfois entrer en concurrence. En cas de règles concurrentes, c'est la règle la plus précise qui sera appliquée en priorité.

Chaque traitement, qu'il soit réalisé par l'automate ou par le gestionnaire, est enregistré dans la table de capitalisation. Il est caractérisé par une origine (gestionnaire, service externe, automate), une portée (pérenne, provisoire), un état de traitement gestionnaire (à traiter, traité, non-traitable, non-traité) et selon les cas par une version de nomenclature de codage.

La portée d'un traitement permet de définir si le traitement doit être capitalisé, c'est-à-dire réutilisé en l'état pour les campagnes suivantes, ou bien si le traitement ne sera pas réutilisé au delà de la campagne en cours.

L'état de traitement gestionnaire fait état du statut de la résolution du cas de reprise, c'est-à-dire qu'il permet de savoir si le cas a été traité ou non-traité par le gestionnaire, puis s'il a été traité, si le gestionnaire a considéré que le cas était traitable ou non-traitable (dans ce cas il fera l'objet d'un redressement automatique).

À l'ouverture d'une campagne, chaque valeur de chaque variable à coder ou à identifier est confrontée à la table de capitalisation, pour déterminer s'il existe une règle de capitalisation qui pourrait s'appliquer. Dans ce cas, le traitement s'interrompt et il n'est pas fait appel aux services externes (Sicore, BRPP, Siam).

2.1.4 Module fonctionnel Référentiel et gestion des nomenclatures

Objectifs du module

Le module fait partie de l'infrastructure du Siera 2. Il reprend les fonctionnalités des applications du Siera 1 qui permettent de traiter les éléments de démographie des entreprises et des établissements, de couverture de champ et de dégroupement des déclarations sociales. Il permet également de gérer de façon centralisée les nomenclatures officielles ou spécifiques utilisées par les processus métier.

Les objectifs principaux du module sont les suivants :

- mettre à disposition des applications clientes un référentiel d'employeurs fondé sur le répertoire statistique des entreprises et des établissements Sirus, avec une historisation des principales variables d'intérêt pour les statistiques d'emploi et de revenus (APE, catégorie juridique, localisation) ;
- prendre en charge les échanges entre les répertoires d'entreprises et d'établissements et le Siera 2 (récupération des unités actives et cessées depuis Sirus, transmission des unités employeuses et des effectifs structurels à Sirus) ;
- gérer l'appartenance des unités employeuses à des champs spécifiques au Siera (champ fonction publique, champ traité par l'Insee dans les statistiques conjoncturelles dans le cadre de la coproduction) ;
- permettre aux équipes métier d'ajouter, modifier ou invalider des clés de dégroupement ;
- ajouter ou modifier des nomenclatures, des tables d'agrégation et des tables de passage.

Le module ne sera pas intégralement développé pour le lot 1. Le périmètre du lot 1 comprend les deux premiers objectifs, ainsi que la gestion des nomenclatures. Les autres composants du module feront l'objet d'une description dans l'étude préalable du lot 2.

Flux d'information

Le sous-module référentiel s'alimente auprès de trois sources :

- en majeur, c'est le répertoire statistique Sirus qui fournit la principale source d'alimentation du référentiel d'employeurs. À intervalles réguliers, Sirus mettra à disposition du Siera 2 l'ensemble des unités actives et des unités cessées à la date d'extraction ;
- en mineur, le référentiel est alimenté :
  - par le résultat des appels Siam générés dans le cadre du processus d'identification du module de traitements mensuels. Ceci permet au référentiel d'être complété entre deux livraisons de Sirus sur le champ des unités nouvellement créées ;
  - par des saisies ad hoc réalisées par les gestionnaires dans le cadre des reprises du module de traitements élémentaires (par exemple pour réactiver une unité cessée statistiquement au sens de Sirus).

Le sous-module gestion des nomenclatures est alimenté directement via l'interactif par les utilisateurs.

Acteurs

Dans la version développée par le lot 1, le sous-module référentiel se limite à des traitements batchs sur lesquels les utilisateurs n'interviennent pas ou très indirectement (au moment de l'identification, en cas d'ajout manuel d'une unité par un gestionnaire).
La gestion des nomenclatures sera prise en charge via une IHM.

Les traitements

Le module développé dans le cadre du lot 1 se limite à trois sous-processus.

Le sous-processus « mettre à jour le référentiel » est commun aux trois types de mises à jour (Sirus, Siam, gestionnaire). Il repose sur la comparaison de l'unité à créer ou à modifier avec les unités du référentiel. Si l'unité est déjà créée, le batch historise les dernières caractéristiques déclarées dans le référentiel et les substitue par les nouvelles données déclarées. Les trois types de mises à jour font l'objet de règles de priorité pour assurer la cohérence entre Sirus et le référentiel sans pour autant empêcher les équipes métier de le remettre en cause. Lors d'une mise à jour du référentiel, les données provenant de Sirus sont prioritaires sur celles déclarées par les gestionnaires, elles-mêmes étant prioritaires sur celles récupérées des appels Siam.

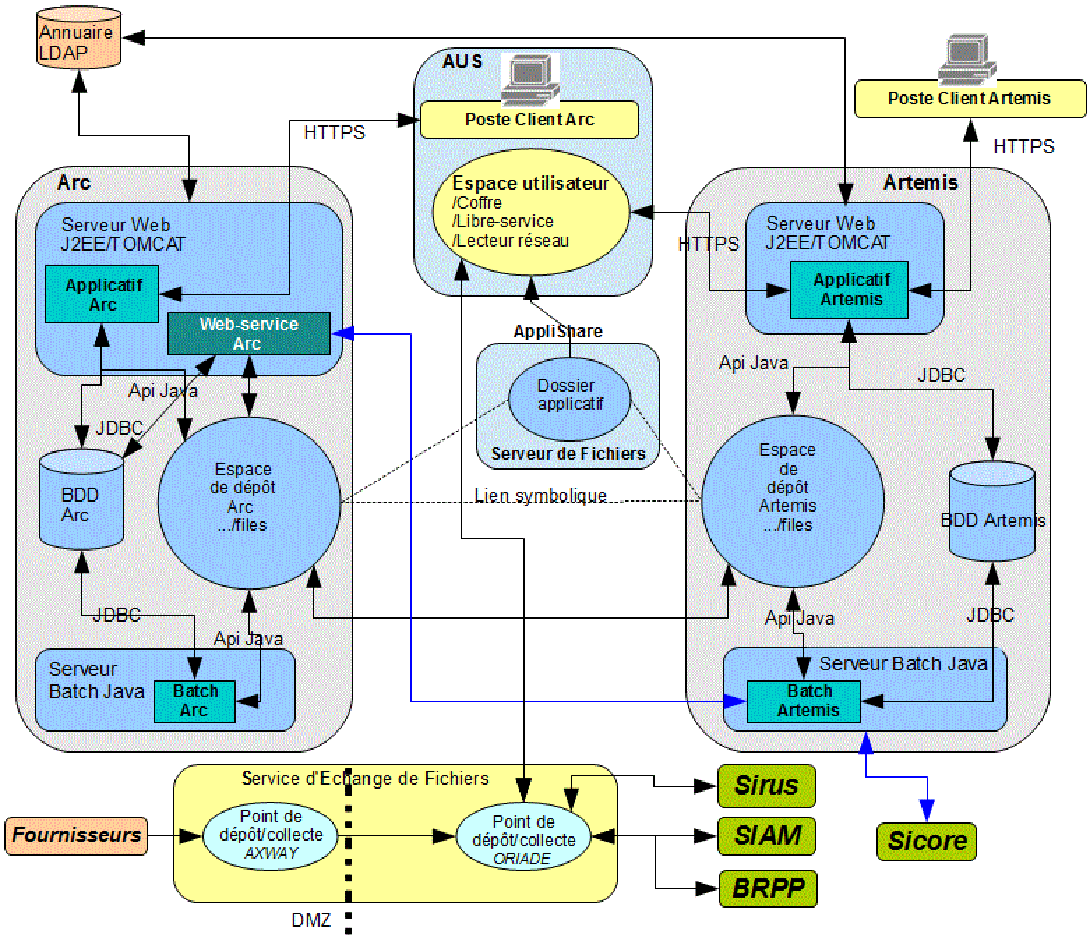
Le sous-processus « vérifier la présence d'une unité dans le référentiel » est un service offert par le module référentiel à destination des applications clientes. Il s'agit d'une forme simplifiée de Siam-identification (on ne vérifie que si l'unité est présente et active dans le référentiel). L'application cliente émet la liste des unités à rechercher, et le module référentiel renvoie pour chacune d'elle si elle fait partie du référentiel, et si oui, son statut (actif ou cessé) et ses principales caractéristiques.

Le sous-processus « gérer une nomenclature » permet à l'utilisateur d'importer une nouvelle nomenclature ou une nouvelle table de passage, ou de modifier des objets existants.

2.2 Architecture matérielle et logicielle

2.2.1 Architecture applicative générale

L'organisation des deux applications ARC et Artemis conçues dans le cadre du lot 1 du projet est traduite dans le schéma ci-dessous.



Ce schéma présente l'architecture applicative définie à la suite des réunions de macro-analyse avec la cellule Architecture et le CEI.

Cette architecture est conforme au schéma directeur informatique et à l'infrastructure du CEI.

Les deux applications ARC et Artemis sont déployées dans le domaine "SIERA" selon une architecture Client/serveur Web, chacune sur une plateforme de production spécifique composée d'un serveur de base de données PostgreSQL 9.4, encodage UTF8, d'un serveur batch Linux-Java 7 et d'un serveur applicatif Web Java 7/Apache/Tomcat 7. Le système d'exploitation est Linux Debian 6.

Les *framework* utilisés pour le développement des interfaces Web J2EE sont *Struts2* et *Spring*.

Par ailleurs, conformément aux préconisations du DPII, un certain nombre de composants sont présents dans les développements des applications du Siera 2 :

composant *InseeConfig* pour la gestion de configuration ;

API de journalisation Log4J pour le gestion des traces applicatives ;

Framework JUnit, DBUnit et *Mockito* pour la mise en place de tests unitaires et fonctionnels automatisables ;

Maven pour la gestion du cycle de vie du projet et la gestion des dépendances ;

la forge f*usionForge* et le logiciel *SVN* comme outil de gestion de version applicative via un protocole de communication sécurisé *SSH* ;

l'environnement de développement Java intégré *Eclipse* ;

Framework Hibernate, pour les correspondances entre les objets Java et la base de données relationnelle, de l'application Artemis.

Enfin, la gestion des authentifications et des profils utilisateurs se fait au moyen de l'annuaire *Ldap*.

Conformément au cadre de cohérence technique, l'indépendance entre les applications est assurée :

- l'accès aux fichiers partagés par les applications et les utilisateurs sera possible grâce à l'utilisation d'un serveur de fichiers lié symboliquement aux espaces de dépôt applicatifs physiques (liens symboliques *Unix, AppliShare)* ;
- le développement de *Webservices* remplacera les liens directs entre bases de données (*database links*)

Le *SEF* gèrera les mises à disposition des fichiers entre le partenaire extérieur (la Cnav dans le cadre de la DSN) fournisseur des déclarations mensuelles et l'Insee, mais aussi la transmission des fichiers demandes/réponses entre les applications et les services (BRPP, SIAM).
L'architecture proposée garantit la sécurité et la confidentialité des données qui seront accédées par les interfaces. L'utilisation du poste de pilotage de l'application ARC permettant de télécharger des données nominatives a été particulièrement sécurisée (cf. §2.2.2).

2.2.2 Accès aux données, sécurité et confidentialité

Exigence de sécurité et de confidentialité

Les exigences de confidentialité ont été précisées dans le document d'expression des besoins :

> **« confidentialité :** les données issues des déclarations sociales comportant des informations nominatives ou directement identifiantes au niveau individuel particulièrement sensibles, la garantie de la confidentialité des données tout au long de la chaîne de réception et de traitement est un enjeu majeur. Les travaux techniques avec le GIP-MDS et la Cnav (qui stockera les données individuelles avant de les transmettre aux organismes de protection sociale ou administrations partenaires) vont démarrer en début d'année 2015. Parallèlement, un dossier sera déposé à la Cnil et une convention sera signée avec le GIP-MDS et la Cnav. Une fois à l'Insee, ces données seront classées haute protection et leur accès sera restreint. »

Par ailleurs, le dossier Cnil précisera la durée maximale de conservation des informations nominatives (nom, prénom et NIR).

Les solutions techniques qui vont être mises en œuvre

À différents stades du processus de production (ou de développement), les acteurs du projet ou de la maintenance doivent accéder aux données, que celles-ci soient stockées sous forme de fichiers ou dans les tables d'une base de données.
Les actions réalisées sur ces données peuvent être de la simple consultation pour analyse ou des traitements plus complexes par le batch ou des procédures de correction de déclarations erronées en "self".
Certaines données sont confidentielles et sensibles ; aussi les choix d'architecture, autant techniques qu'applicatifs, doivent permettre d'en assurer la sécurité et la confidentialité. Ces exigences sont adaptées aux environnements sur lesquels sont faits les accès.

Pour respecter cette nécessité de confidentialité, plusieurs solutions seront mises en œuvre :

- quels que soient les environnements, l'accès aux données sera limité aux seules personnes composant les groupes d'autorisation AD définis par le propriétaire de l'application (ou la maîtrise d'ouvrage pendant la phase projet) ;

- l'authentification sera faite en utilisant l'annuaire Insee Ldap ; les habilitations seront données en fonction du profil des acteurs ;
- les contrôles d'accès seront basés sur la mise en place de droits AD et par l'utilisation d'espaces sécurisés ;
- les données relatives aux NIR seront cryptées sur les postes utilisateurs ; les noms et prénoms ne seront pas accessibles par les gestionnaires.

Ceci est complété par des contraintes de sécurité sur les flux (flux entre la Cnav et l'Insee, ou entre zones internes du CEI). Ainsi, les flux entre les zones internes du CEI sont sécurisés : les transactions, téléchargement et requêtes, entre le poste client et le serveur par l'utilisation du protocole sécurisé HTTPs utilisant le système de chiffrement TLS1 assurant notamment confidentialité et intégrité des données.

Plus précisément, par environnement, la déclinaison des contraintes de sécurité/confidentialité est la suivante :

Environnement de Production

- Les données de production ne seront accessibles en lecture/modification/suppression que par les traitements batchs ou l'interface homme machine (IHM) ouverte aux seuls utilisateurs admis à l'utiliser (gestion des droits par Ldap et profils utilisateur) ;
  - o les communications entre le poste client et le serveur sont sécurisées grâce au protocole HTTPs ;
  - o aucun téléchargement de données n'est autorisé sur les postes utilisateurs : les IHM proposant les fonctions d'upload ne seront accessibles que depuis un espace sécurisé d'AUS. Les données sont automatiquement anonymisées et les fichiers zippés. Ces derniers ne pourront être traités que sur cet espace ;
  - o sur l'IHM, l'affichage du NIR est limité aux 5 premiers caractères ;
- Les accès aux espaces de stockage des données en entrée de l'application ARC ou produites par celles-ci (produits, bilans, traces applicatives) seront soumis à autorisation par le propriétaire de l'application (gestion de groupes AD).

Espace de conservation de fichiers

Cet espace contient les fichiers ou les produits réalisés au maximum pendant 5 ans[12] et qui ne sont plus utiles à la chaine applicative. Les utilisateurs habilités pourront y accéder en lecture seule.

Les données conservées le seront soit sur l'environnement de production, soit dans un coffre sécurisé d'AUS. Les accès seront soumis à autorisation par le propriétaire de l'application (gestion de groupes AD) ; elles seront anonymisées (suppression des noms, prénoms et NIR) afin de respecter les contraintes législatives.

Environnement de qualification fonctionnelle et de tests

- Sur ces espaces, les accès aux données et divers fichiers produits seront soumis à autorisation par le propriétaire de l'application (gestion de groupe AD).
- Les accès aux données (fichiers, logs, exécutables) sont sécurisés (SSH ou FTP).
- Les données personnelles contenues dans les fichiers de test y seront systématiquement cryptées.

Environnement de développement

- Les données utilisées par les développeurs sur leur poste devront être systématiquement cryptées.
- La copie des fichiers entre le poste du développeur et les serveurs est sécurisée (connexion SSH via le client SFTP WinSCP).
- Le lancement des exécutions se fera par des commandes Linux (connexion SSH via l'outil Putty) ou par l'interface graphique mise à disposition des développeurs.

[12] Les différentes durées de conservation seront détaillées dans le dossier Cnil.

2.2.3 Recours aux services, référentiels externes, nomenclatures

Référentiels

Le projet s'appuiera sur le référentiel standard du répertoire d'entreprises et d'établissements Sirus produit par la direction des statistiques d'entreprises. Ce référentiel sera livré trois fois par an. Une évolution vers une mise à jour au fil de l'eau via des Avisirus hebdomadaires sera envisagée à moyen terme.

Concernant la gestion de la géographie, l'étude du recours automatique au référentiel Réfigeo est reportée aux lots ultérieurs. La première version du module se limitera au codage des communes via le service de codification des communes Sicore, couplé à une version chargée manuellement du code officiel géographique en vigueur (cf. infra).

Services

Le projet suppose un recours à différents services de l'Insee pour identifier des personnes ou des établissements et pour coder certaines variables.

Les services concernés sont les suivants :

- **Utilisation des appels Siam**

  Ce service sera appelé - dans sa version Siam/Sirene - pour identifier les établissements à partir de leur identifiant Siret lorsqu'ils n'auront pas été identifiés sans ambiguïté dans le référentiel Sirus. L'appel se fera par un traitement batch (traitement automatique par lot). Un mécanisme de capitalisation permettra de limiter les interrogations aux unités nouvelles, ou à celles qui n'auront pas fait l'objet d'une identification concluante lors des campagnes passées. Ces appels seront effectués une fois, en début de campagne, par un traitement batch asynchrone par dépôt via le SEF d'une demande au format XML. Le nombre de requêtes par fichier sera limité à 10 000 conformément à la documentation technique du composant. Le fichier-retour, également en format XML, sera déposé via le SEF sur l'espace dédié à cette opération.

- **Utilisation du service d'identification BRPP (NIR)**

  Ce service sera appelé pour identifier les salariés à partir de leurs données d'état-civil ou de leur numéro NIR associé aux données d'état-civil. L'appel sera réalisé par un traitement batch (traitement automatique par lot). Un mécanisme de capitalisation permettra de limiter les interrogations. Ces appels seront effectués une seule fois par campagne à l'ouverture, par un traitement batch asynchrone par dépôt via le SEF d'une demande au format XML. Le nombre de requêtes par fichier sera limité à 50 000 conformément à la documentation technique du composant. Le fichier-retour, également en format XML, sera déposé via le SEF sur l'espace dédié à cette opération.

Nota bene : l'utilisation du service d'identification BRPP suppose la publication d'un **décret** en Conseil d'État ou d'un arrêté l'autorisant (selon les cas), le projet de décret/arrêté devra, le cas échéant, être joint au dossier Cnil. L'instruction est en cours avec l'unité des affaires juridiques et contentieuses.

- **Utilisation du service de codification Sicore (commune et PCS)**

  Ce service sera appelé pour coder les communes (du lieu de résidence) et les professions (en PCS à 4 chiffres). L'appel sera réalisé par un traitement batch (traitement automatique par lot) de manière asynchrone[13]. Un mécanisme de capitalisation permettra de limiter les interrogations. L'environnement retenu pour la PCS sera celui actuellement utilisé par DADS (dit PCS-ESE). On notera que le service Sicore est déjà disponible au CEI.

[13]Ce service n'étant actuellement pas offert en standard, il sera étudié une solution interne au projet ou une solution de service offert par le CEI.

2.2.4 Volumétrie et performance

La volumétrie des données à accueillir et à traiter mensuellement, et *a fortiori* annuellement, est un des enjeux majeurs du projet tant pour l'exploitation que pour les performances applicatives.

En effet, à terme, toutes les entreprises des secteurs privé et public devraient renseigner chaque mois une DSN pour leurs salariés. Si l'on distingue trois périodes de montée en charge de la DSN selon les dates d'entrée des entreprises dans le nouveau dispositif[14], le nombre de déclarations correspondantes serait le suivant :

Type de déclarations	Validités 2016-2017	Validités 2018-2019	Validités 2020 et au-delà
DSN (mensuelle)	415 000	494 000	1 897 000
DADS (annuelle)	1 335 000	1 403 000	0

Sous l'hypothèse d'une taille des déclarations concordante avec celle des fichiers tests de DSN reçus par l'Insee il y a quelques mois, cela conduirait à terme à un flux mensuel de la Cnav vers l'Insee de 23 Go et nécessiterait une base de données de production pour ARC de 3,5 To et pour Artemis de 3,5 To (cf. en annexe 4 les informations complètes sur la volumétrie) à la cible. Ces volumétries sont parmi les plus élevées que l'Insee ait à gérer.

Des travaux sont en cours avec la cellule Architecture, le CEI et les services Supports informatiques pour mettre en place un dispositif adapté aux besoins.

D'une part, des espaces de stockage de 4 To sont en cours d'installation au CEI pour permettre l'accueil des bases de données du projet. Si cette taille n'est pas celle utilisée de façon standard au centre d'exploitation, elle reste de taille acceptable. Leur gestion en exploitation va faire l'objet d'une attention particulière dans le cadre du projet et des réunions régulières de travail sont organisées avec le CEI et l'EPOI sur ce sujet (cf. Production/exploitation, § 2.2.5).

D'autre part, le support national Infrastructure de production (SNIP) expertise la possibilité de mettre en œuvre une compression sur une base PostgreSQL, ce qui pourrait permettre de réduire d'un facteur 3 la taille des espaces de stockage et ainsi de rester dans les normes standard actuelles du CEI.

Les exigences de performance sont dictées par un calendrier de production mensuel très contraint au vu des données à traiter mensuellement (cf. calendrier de production, paragraphe 2.1.1).

Les équipes statistique et informatique ont apporté un soin particulier à l'analyse pour que la solution proposée par le projet soit compatible avec les enjeux. Par exemple, le nombre de validité en cours de gestion ainsi que le nombre de variables conservées sont limités au strict nécessaire, certaines solutions proposées et adoptées par le projet Harmonica[15] ont été implémentées lorsqu'elles étaient pertinentes pour le projet, les données téléchargées sont systématiquement compressées, etc.

Des premiers tests de performances ont été réalisés et des tests plus importants seront réalisés à l'été 2015 avec les espaces de stockage de 4 Go, offrant donc l'espace nécessaire pour stocker les fichiers de test et une base de données ayant les caractéristiques identiques

[14] Ces périodes sont encore susceptibles d'évoluer, le dispositif cible n'étant pas complètement défini.
[15] Le projet Harmonica est le premier projet à avoir utilisé des bases PostGreSQL au CEI.

de la base de données de production nécessaire en 2020 ; d'autres tests de performance seront réalisés si la possibilité de travailler sur des bases de données compressées se confirmait.

Par ailleurs, il est envisagé un audit du code et de l'organisation de production courant 2015 pour s'assurer que les choix réalisés sont les plus pertinents ou, le cas échéant, procéder à des améliorations lors du lot 2 du projet.

2.2.5 Production/exploitation

La mise en production

La mise en production fera l'objet d'une attention particulière de la part de l'EPOI car il s'agit d'un premier déploiement de projet au CEI effectué par une équipe du CNIO. Les procédures se mettent en place au CEI et, bien que normalisées à l'Insee, présentent quelques différences avec celles pratiquées jusqu'à présent. La connaissance de l'infrastructure de production du CEI n'est pas encore bien partagée par tous les acteurs.
De plus cette opération va se télescoper avec les bascules des applications en maintenance ce qui pourrait entrainer une disponibilité moindre des équipes RIA-P du CEI et occasionner des dérapages dans le planning.
L'EPOI et la MOA prévoient également dès le lancement de la production en février 2016 un pic de charge inhérent à une première exploitation et au gros volume de fichiers à charger. La disponibilité des équipes devra être garantie sur cette période.

Ces facteurs de risques ayant été identifiés suffisamment tôt, un certain nombre de mesures ont été prises dès mars 2015 pour en limiter les effets :

- des réunions préparatoires et de suivi fréquentes avec le CEI ;
- la préparation de la mise en production dès juin 2015 (rappel : la date de 1ère production réelle est prévue en février 2016 pour l'application ARC) ;
- l'ordonnancement des traitements est pris en charge par l'applicatif ce qui permettra d'alléger la charge d'écriture de la chaîne de production par le CEI ;
- la réalisation de tests de performance dans un environnement de pré-production est prévue très tôt également (été 2015). Ces tests permettront aussi de valider la conformité des développements avec les règles de fonctionnement et l'architecture du CEI.

Cette anticipation dans la préparation de la mise en production a d'ores et déjà permis de mobiliser suffisamment tôt tous les acteurs - cellule Architecture, CEI, supports - autour de nouveaux sujets tels que le montage de bases de données d'un volume supérieur à 2 To et les espaces de conservation des fichiers (dossiers en cours d'étude)..

Le planning de production

Le planning doit prendre en compte les périodes de disponibilité de l'application et les temps nécessaires aux tâches de maintenance et d'exploitation réalisées par le service production. Ce sont par exemple la gestion des sauvegardes et le calcul des statistiques sur les bases de données ; elles seront inventoriées et planifiées, si nécessaire, avec le CEI lors de la préparation de la mise en production de l'application et figureront au contrat de service. Au vu des contraintes de disponibilité des applications, la plage réservée aux opérations du service Production pourrait débuter à 20h00.

Exigences de disponibilité

Disponibilité de l'application ARC

Le poste de pilotage de l'application ARC doit être disponible tous les jours ouvrés entre 7h00 et 20h00.
Le batch ARC sera lancé une fois par jour pour tourner en continu jusqu'au début la plage occupée aux tâches de production. À chaque lancement, l'environnement d'exécution sera initialisé avec les derniers jeux de règles validés la veille. L'application doit être impérativement disponible pendant au minimum 5 jours à compter du 18 de chaque mois de façon à permettre le chargement de l'ensemble des fichiers reçus (cf. calendrier de production, § 2.1.1).

Disponibilité de l'application Artemis

Le poste de reprise de l'application Artemis doit être disponible tous les jours ouvrés entre 7h00 et 20h00 (déploiement uniquement en métropole).
Le batch Artemis pourra être exécuté à n'importe quel moment sur demande de l'administrateur d'application (vraisemblablement chaque début de mois mais ce batch ne pourra probablement pas être planifié). Une grande disponibilité des ressources est attendue tout le temps que durera l'exécution du batch.
Les travaux ne sont pas suffisamment avancés à ce stade du projet pour avoir une idée plus précise de ce que seront les temps de traitement. Toutefois, ce batch lance tous les appels aux services de codification et d'identification - dépôt des requêtes, attente et traitement des retours - et on peut estimer à 2 semaines environ sa durée totale. Les interruptions pour réaliser les tâches de maintenance et d'exploitation seront planifiées avec le CEI de façon à assurer aux gestionnaires la disponibilité de l'application attendue pour les travaux de reprise mensuels.

2.3 Les futurs utilisateurs

Deux postes de travail sont prévus dans le lot 1 :

- poste de travail *ARC* il permettra l'accès aux fonctionnalités en interactif des modules fonctionnels « accueil-réception-mise au format » et « gestion des nomenclatures » ;
- poste de travail *Artemis* qui permettra de réaliser les traitements élémentaires automatiques et les reprises gestionnaire.

Les utilisateurs de ces deux postes de travail se situent soit au sein de la division Exploitation des fichiers administratifs pour l'emploi et les revenus (EFA) soit au centre Statistiques sociales et locales de Metz (CSSL).

Le poste ARC sera utilisé par l'administrateur d'application (et les responsables des processus) et le responsable informatique de l'application.

Pour le poste Artemis, les utilisateurs se répartissent en quatre rôles distincts :

- les gestionnaires (CSSL) : ils sont responsables des traitements de reprise d'échecs d'identification des employeurs ou de codage commune ou profession ;
- le chef d'équipe (CSSL) : il est responsable de l'organisation du travail des gestionnaires, en particulier, il gère l'alimentation des portefeuilles des gestionnaires ;
- les experts des processus de capitalisation (CSSL, division EFA) : ils sont responsables du suivi qualité des reprises et de la gestion des tables de capitalisation ;
- l'administrateur d'application et les responsables des processus à Paris et à Metz, sous la responsabilité de la division EFA.

Ces différents rôles correspondent chacun à un profil d'utilisation de l'application et à un poste de travail dédié. Les rôles seront affinés en termes d'activités dans le cadre de la démarche Maiol et le groupe Maiol participera aux travaux de définition des écrans du poste Artemis. Un soin particulier sera porté sur l'ergonomie et le respect des règles d'accessibilité (versions application interne Insee) pour éviter toute exclusion.

2.4 Les coûts de fonctionnement de la solution privilégiée

Le périmètre du lot 1 du projet constitue une première étape insuffisante pour disposer de l'évaluation d'un dispositif intermédiaire par rapport à la cible Siera 2 comparable à un sous-ensemble identifiable de l'existant.

Concernant les coûts gestionnaires (CSSL), un premier exercice a été réalisé pour le CD-PTT de juin 2015 concernant les besoins pour traiter la DSN en 2016 et 2017. Il a été programmé 5 ETP de profil gestionnaire pour la partie traitements élémentaires de la première année. L'évaluation sera affinée en fonction de la montée en charge de la DSN, des volumes de rejets à traiter, de l'efficacité de la capitalisation et des temps de traitement moyen.

Concernant l'encadrement des gestionnaires (CSSL), il est prévu de démarrer avec un préfigurateur et d'attendre les recommandations de la démarche Maiol pour la mise en place de l'organisation intermédiaire (2016-2017) puis cible.

*Concernant les coûts d'administration statistique (*administrateur d'application et responsables des processus), ils devront être estimés en tenant compte à la fois des applications qui passeront en maintenance (application ARC en 2016) et de la charge de paramétrage des contrôles et de mise au format statistique. Concernant l'application Artemis, la deuxième version sera prise en charge par l'équipe de projet statistique y compris pour les maintenances correctives.

Concernant les coûts de maintenance informatique, ils sont généralement entre 10% et 20% du coût de développement. Pour ARC, ils ont été estimés à 60 jours par an, hors travaux d'évolutions fonctionnelles importants ; pour Artemis, ils ont été estimés à 70 jours par an, hors travaux d'évolutions fonctionnelles importants.

Concernant les coûts de production informatique, nous ne disposons pas actuellement de suffisamment de recul sur les coûts de production CEI, avec la mise en place du nouveau data center et l'effort d'automatisation qui a été réalisé.

Concernant le budget (coûts externes), les principes de tarification (participation au coût de mise à disposition des données retenus pour la DSN) devraient être proches de ceux existants pour les DADS (l'Insee n'étant pas concerné par les coefficients de complexité des filtres). Le principe d'un gel des montants 2014 a été proposé pour couvrir la période de concomitance DSN/DADS. Une fois le système entièrement couvert par la DSN, la participation de l'Insee pourrait être revue à la baisse suite à l'augmentation du nombre de partenaires[16].

Les évaluations de *coûts de formation et les frais de déplacements* sont prématurées et doivent être faites sur un périmètre plus large.

[16] Informations données par la Cnav lors de la réunion budgétaire des partenaires TDS de mai 2015, en attente de confirmation.

3 MISE EN ŒUVRE DE LA SOLUTION PRIVILÉGIÉE

3.1 Stratégie de développement

3.1.1 Méthode de projet

Compte tenu du besoin de flexibilité affiché pour les développements des applications du lot 1 de Siera 2, la méthode de projet choisie est en cycle itératif court par prototypage.

Les développements suivent un processus itératif de 3 semaines au bout duquel chaque prototype est évalué avec les utilisateurs et amélioré en fonction des problèmes rencontrés lors des tests. Le système complet se construit progressivement par itérations successives.

Cette méthode a pour avantage de converger rapidement vers une solution opérationnelle. Par contre, un prototype n'est nativement pas qualifié aux normes de développement et de production. L'équipe informatique doit donc planifier régulièrement des mises à niveau qui représentent 2,5 jours de développement par mois.

Actuellement, les méthodes agiles sont en expérimentation à l'Insee pour quatre projets, dans leurs phases de développement, après l'étude préalable. Après discussion, il n'y a pas eu consensus des acteurs pour démarrer le lot 1 en méthodes agiles. Or l'adhésion de tous les acteurs à ces méthodes est un facteur indispensable de réussite. Il a semblé qu'il n'était pas opportun de démarrer le projet en Agile pour trois raisons[17] :

- les délais pour le lot 1 du projet nécessitent de conduire conception et développement en parallèle. Cela supposait donc une expérimentation nouvelle différente de celles en cours à l'Insee ;
- le coût d'entrée dans les méthodes agiles est important (de l'ordre de quelques mois avant de trouver un mode de fonctionnement stabilisé). Cela semblait difficilement compatible avec les délais serrés du lot 1 ;
- la collaboration étroite avec les équipes travaillant sur les applications en production pouvait être difficile à mettre en œuvre entre une équipe projet fonctionnant en Agile et les équipes ayant des contraintes de disponibilités liées aux phases de production de l'application.

3.1.2 Modularité des développements informatiques

Dans un contexte de mise en production par lot des nouvelles applications ARC et Artemis et de leur intégration dans un système existant, il semble essentiel de limiter l'impact des inadéquations possibles entre les choix effectués pour le lot 1 et les besoins des projets ultérieurs. Les développements du lot 1 seront donc réalisés en tenant compte de la cible telle qu'envisagée par la maîtrise d'ouvrage. Ils ont également été conçus pour être facilement remplaçables, voir jetables, avec une granularité fonctionnelle assez faible pour éviter d'avoir à reprogrammer des pans entiers du système cible.

D'une façon plus générale, la stratégie de développement doit permettre de gérer le manque de visibilité sur le contenu final de certaines fonctionnalités, de rendre peu impactant l'indisponibilité éventuelle de certains services externes, ou encore de rendre possible la priorisation sur le champ de fonctionnalités en cas de retard, besoin déjà exprimé dans l'expression des besoins.

Dans un souci de cohérence globale du système d'information, il est par ailleurs souhaitable que les fonctionnalités développées dans le cadre du programme d'évolution du Siera soient mutualisables pour d'autres applications du Siera et que leur code soit centralisé pour que leurs évolutions s'effectuent de façon homogène sur l'ensemble du système d'information.

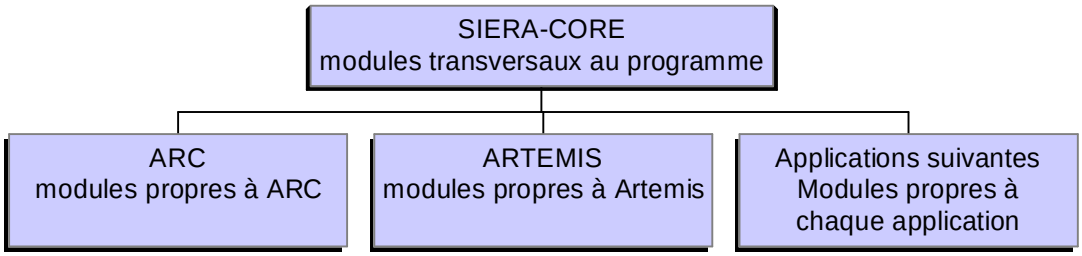
[17] Le DAP ne préconise l'utilisation de l'agilité que lorsque tous les membres de l'équipe projet sont volontaires pour utiliser cette méthodes. Ce n'était pas le cas de l'équipe projet de Pirénés.

Pour répondre à ces besoins, il est proposé de développer le lot 1 en briques de fonctionnalités élémentaires appelées "modules[18]".

À la manière d'un service, chaque module interagit avec le système cible via deux interfaces uniques d'entrée et de sortie. Chaque module crée son propre contexte d'exécution et est ainsi désolidarisé des autres modules et de son environnement d'exécution.

Grâce aux technologies java, le surcoût du développement par module est faible, l'essentiel étant de spécifier la granularité des différents modules.

Architecture de développement modulaire



Avantages de la modularité

- est adaptée aux méthodes itératives de conduite de projet ;
- apporte de la cohérence au système dans son ensemble ;
- rend très flexibles les développements et la maintenance ;
- permet de porter nativement les programmes sur différents types d'interface et facilite l'implémentation des tests fonctionnels ;
- possibilité pour la maîtrise d'ouvrage pour les applications ARC et Artemis d'essayer les différents modules tels qu'ils seront implémentés en production (« bac à sable ») ;
- entraîne un faible coût de développement (environ 15 jours sur l'ensemble du lot 1 du projet Pirénés).

Inconvénients de la modularité

Cela oblige à découper les fonctionnalités avec une granularité fine et indépendantes. Ceci n'est pas toujours possible pour des fonctionnalités complexes et imbriquées. Mais les conceptions générales des applications ARC et Artemis ont été élaborées avec cet objectif de simplicité.

3.1.3 Généricité

Certaines fonctionnalités sont développées de façon à pouvoir adapter un même processus à des objets différents, soit automatiquement, soit à travers une interface de paramétrage. Ces développements sont alors qualifiés de génériques.

Le coût de développement des fonctionnalités génériques est très élevé et dépend directement du niveau de généricité souhaité. Il est donc essentiel d'évaluer les gains attendus par rapport à l'investissement nécessaire.

En outre, les fonctionnalités génériques paramétrables par l'utilisateur peuvent créer des effets non désirés sur une chaîne de production. Il est donc aussi nécessaire de normaliser l'insertion de nouveaux paramètres, et de proposer aux utilisateurs et au RIA la possibilité de simuler et valider leurs effets avant leur passage en production.

Implémentation de la généricité dans le lot 1

L'application ARC, permettant de paramétrer via une IHM, le chargement, le normage, le contrôle, le filtrage et la mise au format statistique de différents type de fichier doit donc remplir ces critères :

[18] Il s'agit ici de modules informatiques, différents des modules fonctionnels décrits dans la partie précédente.

- d'une part, les normes d'échange des fichiers fournis par la Cnav évoluent rapidement. À titre d'exemple, le passage à la norme N4DS a coûté 340 jours de maintenance adaptative au Siera (cf. bilan[19] du projet N4DS lot 1). La décision de créer l'application ARC générique se justifie donc car le coût de développement de cette application est également estimé à 300 jours.
- d'autre part, le développement modulaire de l'application ARC permet aux statisticiens et au RIA de tester dans un « bac à sable » avant de statuer sur une mise en production.

Chacune des phases de chargement, de normage, de contrôle, de filtrage et de mise au format statistique fait l'objet d'un paramétrage par l'utilisateur de l'application : la façon de détecter la version de la norme et la validité, les jeux de contrôles, les règles d'exclusion ou encore les règles de « mapping » sont autant de paramètres qui sont spécifiés via l'interactif par les utilisateurs. La combinaison de ces règles crée un contexte dans lequel vont s'exécuter les différents traitements. En ce sens l'application ARC est totalement générique, puisque les paramètres sont saisis par les utilisateurs et déterminent totalement l'ensemble des traitements subis par les fichiers déposés.

Il est envisagé d'implémenter d'autres fonctionnalités génériques dans le lot 1 mais à l'inverse de ce qui est proposé dans l'application ARC, ces dernières ne seront pas dynamiquement configurables par l'utilisateur :

- un composant de développement rapide d'IHM : les écrans développés dans le lot 1 utilisent une même fonctionnalité générique ;
- la gestion centralisée dans l'application ARC des tables de nomenclatures : toutes les tables de nomenclature respectent un formalisme identique ;
- les robots d'identification utilisés dans l'application Artemis (NIR, Siret, commune de résidence, PCS)
    - leur fonctionnement décisionnel est spécifié par les statisticiens grâce à des tables de vérité (tableaux de Karnaugh) ;
    - le RIA remplace ces tables de vérité dans le code de l'application Artemis pour modifier le comportement des robots et doit éventuellement développer les nouveaux scenarii de traitement s'ils n'existent pas déjà.

3.1.4 Stratégie de tests

Des tests de plusieurs natures seront menés sur le projet. Si ces tests sont classiquement réalisés pour tout projet d'une certaine ampleur, l'accent est mis particulièrement sur deux aspects :

- la mise en place d'un outil de test qui permet de réaliser les tests fonctionnels, partagés par l'EPOI et l'EPS, et de s'assurer de la non-régression lors de corrections ou d'ajouts de fonctionnalités ;
- la réalisation de tests de performance. L'enjeu est important pour le projet compte tenu des volumétries des flux et des bases.

La mise en place d'un outil de test

Dans le cadre du projet Pirénés, un outil interne au CNIO[20] a été mis en place pour faciliter la mise en place et l'historisation des tests, ainsi que leur exécution et l'analyse des résultats produits.

L'utilisation de cet outil aura vocation a être étendue, dans un premier temps, à l'ensemble des applications du programme d'évolution du Siera. Quelques applications existantes du programme Siera pourront également en bénéficier pour réaliser leurs tests de qualification avant la bascule au CEI.

[19] Note n° 1639/DG75-F201 du 27 juillet 2012
[20] Cet outil a été présenté à la cellule Architecture et au SGI. Il est conforme au cadre de cohérence technique de l'Insee.

La mise au point des jeux de règles (bac à sable)

Le poste de pilotage de l'application ARC proposera au statisticien (l'EPS puis l'administrateur d'application), dans l'environnement de production, des fonctions de mise au point des règles de contrôles, de filtrage et de transcodage en variables statistiques des variables administratives véhiculées par les déclarations sociales. Le statisticien spécifiera les règles à l'aide d'un pseudo-langage à la syntaxe aussi simple que possible, proche du SQL. Le nombre de types de règles possible est par ailleurs limité. Ces règles sont ensuite traduites par l'outil en langage informatique exécutable (des requêtes SQL).

Un mécanisme permettra de tester syntaxiquement la validité de la requête formée. Une fois cette étape franchie, le statisticien décidera le déploiement en production du jeu de règles ainsi validé.

Les tests de performance

La double problématique de la volumétrie et du calendrier mensuel de production impose de s'assurer de la performance de l'application ARC. Elle devra en effet pouvoir charger et contrôler 23 Go (compressés) de données en six jours à compter du 18 de chaque mois.

De plus, ces volumes pourront se cumuler chaque mois, puisque les bases de données de ARC ne seront purgées qu'après le transfert des données aux applications clientes.

Des premiers tests ont eu lieu avec les données de la phase 1 de la DSN et sont rassurants sur les performances obtenues. Une prestation est prévue pour auditer le code et le paramétrage de l'infrastructure et préconiser d'éventuelles d'optimisations à l'automne 2015.

Les tests de charge

Les tests de charge pour les deux postes de travail développés dans le lot 1 seront menés en utilisant l'outil JMeter avec l'appui de la ressource test du CNIO. Ils permettront de valider le bon fonctionnement de l'application en pleine charge : 60 utilisateurs minimum simultanés utilisant les fonctions les plus consommatrices de ressources. Les scénarios éligibles aux tests seront décrits et priorisés dans un document dédié au plan de test. Un rapport de test sera livré. Tous les éléments utiles à la réalisation des tests étant conservés et historisés grâce à l'outil de test, les campagnes de test pourront être renouvelées autant de fois que nécessaire.

3.1.5 Mode de fonctionnement entre les équipes

Les spécifications et les relations entre les équipes projet

Compte tenu de la taille des équipes projet, une organisation spécifique a été mise en place pour la conception, la spécification et la recette des développements.

La phase de conception est du ressort des deux chefs de projet. Les documents de conception générale sont réalisés sur le périmètre de chaque module fonctionnel, puis débattus lors de réunions réunissant l'ensemble des équipes informatique et statistique. La conception repose sur l'élaboration des diagrammes UML de cas d'utilisation et des diagrammes UML de classes.

Les spécifications détaillées et les développements informatiques de chaque cas d'utilisation sont confiés à un binôme de l'EPS/EPOI clairement identifié. Les spécifications et développements sont validés par les chefs de projet statistique et informatique. Autant que possible, les cas et jeux de tests sont élaborés dès la phase de spécifications.

La spécification de toutes les IHM est confiée à un unique CPS, qui participe par ailleurs à la démarche Maiol. Chaque écran des postes des gestionnaires est discuté au sein du groupe Maiol afin d'assurer une bonne prise en compte des besoins des utilisateurs. Le chef de projet informatique en charge du développement des IHM participe également en tant qu'invité à ces discussions.

Les recettes ne sont pas uniquement réalisées par le CPS qui a spécifié le cas d'utilisation. Ceci permet de mieux valider les spécifications, de rendre l'organisation de l'équipe statistique plus souple et d'assurer la cohérence de l'ensemble.

La migration des données

Le lot 1 ne donne pas lieu à une migration de données issues des applications du Siera existant.

La documentation

La documentation du projet est disponible sur Gforge à l'adresse suivante :

http://gforge.insee.fr/projects/dsn/

Ce site est géré et mis à jour par l'EPOI. Il contient des éléments sur les événements du projet, ses acteurs, la documentation technique et les spécifications. Une documentation métier sera également mise à disposition du propriétaire d'application en fin de projet. Une vigilance spécifique sera apportée à la documentation métier de l'application ARC, en particulier sur la syntaxe et l'écriture des règles de contrôle et de mapping.

Par ailleurs, les principales notes du programme d'évolution sont présentes sur le site intranet de la maitrise d'ouvrage, disponible à l'adresse suivante :

http://www.agora.insee.fr/jahia/Jahia/site/dg-dsds/pid/145616

3.2 La démarche Maiol

Une démarche Maiol a été mise en place[21]. Le groupe Maiol est composé du Reso (Albane Gourdol, CSSL), des responsables des équipes DADS (Annie Herbin, DR de Champagne-Ardenne et Odile Thirion, DR de Bourgogne), d'une représentante de la division EFA (Christine Couderc), d'une des CPS de l'équipe projet (Adeline Baudrey) et du conseiller interne en organisation de la division Camap (Monique Bourbigot-Pognat).

La réunion de lancement de la démarche a eu lieu le 10 février 2015. L'analyse de l'existant sera finalisée pour début juin. Les prochaines échéances concernent les travaux sur l'IHM du poste Artemis (juin/juillet 2015) et des propositions d'organisations cible pour l'arrivée de la DSN en fin d'année 2015.

3.3 Planification

Le projet Pirénés est composé de deux lots présentés dans l'expression des besoins et rappelés en début du présent document.

Le lot 1, objet de la présente étude préalable, est lui-même découpé en sous-lots correspondant chacun à une application :

- l'application ARC correspondant à l'accueil, la réception, le contrôle et la mise au format statistique des données DSN, ainsi que la gestion des nomenclatures ;
- l'application Artemis de traitements statistiques élémentaires des données, automatisés ou manuel, ainsi que la gestion du référentiel d'employeurs (version 1).

L'application ARC sera développée et testée pour entrer en production à la fin 2015 pour une première campagne réelle en février 2016.

L'application Artemis sera développée et testée pour une mise en production d'une première version en mars 2016, c'est-à-dire pour permettre le traitement des premiers fichiers DSN reçus courant février 2016.

Compte tenu du volume de déclarations à traiter chaque mois à compter de février 2016, la maîtrise d'ouvrage souhaite, en cas de dérapage du calendrier de développement, privilégier le maintien de l'échéance de mars à celui du périmètre fonctionnel. En conséquence, le comité de pilotage de mai 2015 a d'ores et déjà priorisé les développements de l'application Artemis pour tenir compte de ce besoin. Ainsi les fonctionnalités indispensables aux traitements manuels seront développées en priorité dès l'été 2015. Une partie des

[21] cf. note n°318/DG75-F201 et 113/DG57C003 du 19 février 2015.

fonctionnalités restantes (poste de pilotage, clôture de campagne) ont été identifiées comme moins critiques et pourront, en cas de besoin, être décalées jusqu'à l'été 2016.

Les grandes étapes de la planification générale du projet sont donc les suivantes :

<table>
<tr><th>Année</th><th>Mois</th><th>Application ARC</th><th>Application Artemis (V1)</th><th>Jalons juridiques</th><th>Jalons organisationnels du projet</th><th>Maiol</th></tr>
<tr><td>2</td><td>09</td><td rowspan="3">Conception générale</td><td rowspan="4"></td><td></td><td></td><td></td></tr>
<tr><td>0</td><td>10</td><td></td><td rowspan="3">Rédaction de l'expression des besoins</td><td></td></tr>
<tr><td>1</td><td>11</td><td></td><td></td></tr>
<tr><td>4</td><td>12</td><td rowspan="4">Développement et tests prototype V1</td><td></td><td></td></tr>
<tr><td rowspan="11"></td><td>01</td><td rowspan="4">Conception générale</td><td></td><td>Séminaire de lancement du projet</td><td></td></tr>
<tr><td>02</td><td></td><td rowspan="2">Rédaction de l'étude préalable
Réunions de macro-analyse</td><td>. Note de périmètre
. Première réunion du groupe</td></tr>
<tr><td>03</td><td></td><td rowspan="3">Analyse de l'existant</td></tr>
<tr><td>2</td><td>04</td><td>Développement et tests prototype V2</td><td></td></tr>
<tr><td>0</td><td>05</td><td rowspan="4">Développements et tests:
accès aux services externes
IHM gestionnaires
ouverture de campagne
automates</td><td></td></tr>
<tr><td>1</td><td>06</td><td rowspan="2">Tests de performance</td><td></td><td></td></tr>
<tr><td>5</td><td>07</td><td>Dossier Cnil[22]et projet de décret en CE pour l'utilisation du NIR si besoin</td><td></td><td></td></tr>
<tr><td>08</td><td></td><td></td><td></td><td></td></tr>
<tr><td>09</td><td></td><td>Projet de conventions avec la Cnav et le GIP-MDS</td><td></td><td></td></tr>
<tr><td>10</td><td rowspan="2">Pilote phase 3</td><td rowspan="3">Développements et tests :
IHM pilotage
Clôture de campagne</td><td></td><td rowspan="4">Rédaction de l'étude préalable du lot 2</td><td></td></tr>
<tr><td>11</td><td></td><td></td></tr>
<tr><td>12</td><td>Mise en production</td><td></td><td>Rapport intermédiaire sur les organisations cible</td></tr>
<tr><td rowspan="2">2</td><td>01</td><td></td><td rowspan="2">Tests de performance</td><td></td><td></td></tr>
<tr><td>0</td><td>02</td><td>Première campagne</td><td></td><td></td></tr>
<tr><td>1</td><td>03</td><td></td><td>Mise en production
Première campagne</td><td></td><td></td><td></td></tr>
</table>

[22] Le dossier de l'Insee doit être articulé avec celui de la direction de la sécurité sociale sur la DSN et l'arrêté filtre des données transmises à l'Insee.

3.4 Budget de la solution

Méthode d'estimation des charges

Les charges du lot 1 du projet ont été estimées en adaptant l'estimation réalisée par la société Semantys pour le projet Siasp-ONP à travers des interviews régulières de l'équipe projet.

Le projet Siasp-ONP devait initialement réaliser l'accueil-contrôle-transformation en données statistiques des fichiers administratifs pour le compte du programme d'évolution du Siera. L'abandon du projet SI-Paye de l'ONP par le gouvernement ayant entrainé la suspension du projet Siasp-ONP en juillet 2014, ce module fonctionnel a finalement été pris en charge par le projet « Pirénés - lot 1 ». Par ailleurs, plusieurs autres fonctionnalités de Siasp-ONP sont également présentes dans le projet « Pirénés - lot 1 », comme par exemple la constitution d'un référentiel d'employeurs.
Par anticipation, il avait été demandé à Semantys d'isoler dans le rapport définitif d'estimation des charges de Siasp-ONP le périmètre commun avec le futur projet. Ainsi, les équipes projet se sont appuyées sur les documents déjà produits pour les faire évoluer afin de prendre en compte les particularités nouvelles du nouveau projet.

Parallèlement, le CPOI a réalisé une estimation des charges pour l'EPOI *à dires d'expert*. Les deux estimations convergent à 7% près, venant conforter l'estimation.

Le chiffrage des coûts du lot 1

Le chiffrage des coûts de la solution commencent à partir de la phase de conception du projet, c'est-à-dire à compter de septembre 2014. Il se fonde sur une fin de projet en mars 2016. Le coût du projet inclut donc les travaux réalisés lors de l'étude préalable, tels que macroanalyse, prototypage, conception de la solution technique, etc. Il inclut également les travaux du Maiol dans sa première phase (jusqu'à la fin d'année 2015).

En revanche, il ne comprend pas les travaux qui seront réalisés par l'équipe statistique à partir du second semestre 2015 de conception du lot 2, dont l'étude préalable devrait être remise au premier trimestre 2015.

Estimation des charges du projet « Accueil et traitements de la DSN - lot 1 »[23]

	Charges (en jours)	Équivalent ETP (jour/homme)	*Dont équivalent ETP jusqu'à la fin de l'étude préalable*
MOA/MOAD	240	270	*110*
EPS	900	1 000	*520*
EPOI	1 220	1 360	*650*
Autres	730	810	*480*
Total	**3 090**	3 430	*1 760*

Les ressources informatiques et statistiques du projet sont les suivantes :

- équipe informatique (EPOI) :
  - 2 CPOI (arrivés en septembre 2014)
  - 2 analystes (arrivés en octobre et en novembre 2014)
  - 1 développeur (arrivé en juillet 2015)

  soit 1 280 jours ETP sur l'ensemble de projet dont 620 jours ETP jusqu'à la fin de l'étude préalable.

[23] L'estimation tient compte également de la participation de certains acteurs (MOA, responsable statistique, informatique et Reso le cas échéant) aux comités transversaux du programme d'évolution du Siera.

- équipe statistique (EPS) :
    - 1 responsable statistique (arrivé en octobre 2013)[24]
    - 4 chefs de projets statistiques et chefs de projets statistiques adjoints (arrivés entre septembre et décembre 2014)

  soit 1 380 jours ETP sur l'ensemble du projet dont 750 jours ETP jusqu'à la fin de l'étude préalable.

Budget total du lot 1

Le budget de la solution, présenté en tableau ci-après, est estimé sur la base des travaux réalisés avec la société Semantys, réévalués par l'équipe projet sur le périmètre fonctionnel restreint (*cf. supra* « Méthode d'estimation des charges »).

Budget total de la solution

	De septembre 2014 à mars 2016	
	En jours	En Euros[25]
Coûts de développement informatique	1220	464 000
Coûts de maîtrise d'ouvrage	240	115 000
Coûts de maîtrise d'œuvre statistique	900	358 000
Coûts de production informatique	100	28 000
Coûts autres acteurs internes (utilisateurs, cellule Architecture, Camap, Réso, etc.)	560	190 000
Coûts de formation		_
Coûts de déplacements		5000
Coût prestations externes		140 000
Total	3 020	

[24] Une partie de travaux réalisés par le responsable relève de la maitrise d'ouvrage déléguée. Non compris les travaux sur le projet Siasp-ONP. Les ressources ne sont comptées qu'à partir de septembre 2014.

[25] La valorisation de travaux a été effectuée avec les coûts de base de la note n°106/DG75-C101/PG/NB du 6 juin 2014, soit 398 euros pour un cadre A et 282 euros pour un cadre B.

3.5 Analyse de risques de la solution privilégiée

Pendant le séminaire de lancement de projet, une première analyse des risques sur le programme a été menée et a permis d'identifier 6 catégories de risques présentés dans le dossier d'expression des besoins :

- risques liés au projet externe DSN ;
- risques liés aux ruptures statistiques ;
- risques liés au cycle de production des données d'emploi ;
- risques techniques ;
- risques organisationnels ;
- risques liés à la stratégie de développement.

Lors de la deuxième réunion du comité directeur du programme d'évolution du Siera, la maitrise d'ouvrage a proposé un classement des risques en fonction des structures en charge du suivi (comité directeur ou comité opérationnel du programme) :

- les risques liés aux calendriers des trois investissements du programme ;
- les risques liés à la qualité des données (en phase de montée en charge ou en différentiel avec les dispositifs actuels) ;
- les risques liés à la transition.

On notera sur ces risques que la montée en charge de la DSN vient de passer deux paliers significatifs : la mise en production de la phase 2 (sur la partie substitution des BRC) et l'obligation intermédiaire pour les employeurs visés par le décret de septembre 2014[26]. Concernant la qualité des données, suite aux remontées des premiers utilisateurs (Cnam, Pôle emploi, Dares et Acoss), des ateliers pour améliorer la qualité des déclarations ont été mis en place avec les éditeurs (cf. fiche de suivi du programme d'évolution du Siera en annexe).

Les autres risques doivent être suivis au niveau des projets. Il s'agit plus particulièrement des risques techniques, des risques organisationnels (liés à l'acceptation des postes de travail, à la taille et à la durée du programme) et les risques liés à la stratégie de développement. La description suivante se limite au périmètre du projet Pirénés.

Risques techniques

- *Performance et volumétrie des données*

La volumétrie des données reçues sera très importante, et pourrait représenter jusqu'à douze fois la volumétrie des DADS en nombre de lignes salariés. Une vigilance toute particulière devra donc être apportée à la robustesse de la procédure de chargement, mais également aux performances des applications qui seront utilisées quotidiennement par les gestionnaires.

Facteur : volumétrie et rythme des données reçues et traitées

Impacts : problèmes de performance, difficultés à traiter les données sur un rythme mensuel

Parades : informaticiens expérimentés dans l'équipe informatique, programmation des tests de performance dès 2015 (avec la volumétrie cible) (cf. § 3.1.4), travail en commun avec le CEI plusieurs mois avant la mise en production.

Probabilité[27] : ++ Gravité : +++

[26] Décret n° 2014-1082 du 24 septembre 2014 portant obligation pour les employeurs payant plus de 2 millions d'euros de cotisations ou 1 million d'euros via un tiers-déclarant déclarant plus de 10 millions d'euros de cotisations.
[27] Les échelles utilisées vont de + (faible), ++ (moyen), +++ (fort)

- *Transfert des applications (Siera et services utilisés) au CEI*

Le développement et la mise en production du lot 1 sont concomitants au transfert des applications du Siera et de certains services utilisés vers le CEI.

Facteur : évolution de la production informatique de l'Insee concomitante à la montée en charge du CEI

Impacts : charge du CEI, des équipes métiers et informatiques, évolutions à prendre en compte pendant les phases de déploiement

Parades : suivi des calendriers, réunions régulières avec le CEI

Probabilité : ++ Gravité: ++

Risques organisationnels

- *Acceptation du nouveau poste de travail Artemis*

Si la réorganisation des travaux et les risques associés seront suivis par le comité de suivi du centre de Metz et le CD-PTT, le projet devra être vigilant sur l'acceptation par les gestionnaires du nouveau de poste de travail et en particulier de sa première version compte-tenu des délais.

Facteur : poste inachevé dans les délais

Impacts : décalage du démarrage des travaux des gestionnaires, mécontentement

Parades : implication de la démarche Maiol dans la définition du poste de travail, priorisation des travaux

Probabilité : + Gravité: ++

- *Gestion des échéances du programme*

La taille du programme et les délais de réalisation très contraints du premier projet imposent aux équipes statistique et informatique de gérer des investissements aux horizons temporels différents : le premier projet de court terme indispensable à la prise en compte de la DSN, les études indispensables à la détermination du périmètre des projets à venir.

Facteur : gestion de plusieurs échéances à l'échelle des projets de réingénierie

Impacts : risques de dérapage des projets ultérieurs

Parades : programmation des jalons, anticipation dans les programmes de travail

Probabilité : + Gravité: +

- *Gestion de la mobilité au sein des équipes*

Le troisième risque organisationnel à gérer au niveau des projets de réingénierie est lié à la durée prévue du programme et au cycle de mobilités des équipes. L'organisation du programme en plusieurs projets implique une durée totale de l'investissement d'au moins 6 années pour couvrir l'ensemble du périmètre fonctionnel envisagé. Dans ce contexte, il est très vraisemblable que les équipes constituées en septembre 2014 ne seront pas identiques à celles qui achèveront le dernier projet du programme.

Facteur : mobilité dans les équipes projet, de la maîtrise d'ouvrage

Impacts : perte de la connaissance, changement de cible

Parades : anticipation, découpage en projets compatibles avec les mobilités, documentation, transfert de compétences lors des passations de postes.

Probabilité : +++ Gravité: +

Risques liés à la stratégie de développement

- *Découpage du système cible en modules*

Le développement par module permet de désolidariser les développements des modules les uns des autres. Cela permet d'avoir un développement flexible, ce qui répond mieux aux contraintes de délais et au manque de visibilité sur certaines phases du processus cible et de connaissance sur les données futures en entrée de la chaîne.

Facteur : travail en silo sur la spécification et le développement des modules

Impacts : incohérences entre les différents modules, difficultés lors des raccordements

Parades : veiller tout au long du projet à l'analyse du système d'information dans son ensemble, relecture commune des conceptions générales, participation de l'ensemble des acteurs aux recettes, mise en place systématique de tests de requalification de l'ensemble de la chaîne applicative à la mise en production de chaque module

Probabilité : + Gravité: +

4 LA SUITE …

4.1 Projet Pirénés (lot 2)

L'objectif de ce deuxième lot est de permettre la constitution du produit *Tous salariés* intégrant les données issues des processus Siasp (emploi public) et PE (salariés des particuliers employeurs) et du nouveau processus pour les déclarations DSN ou DADS (secteur privé).

Le périmètre du projet reste inchangé par rapport à l'expression des besoins :

- version minimale du module de traitements statistiques structurels (traitement de la cohérence emploi/salaire, constitution du produit « Tous salariés » et traitement de l'exhaustivité) ;
- composante « traitement des dégroupements » du module référentiel.

Le calendrier prévisionnel du second lot est :

Étude préalable : remise au premier trimestre 2016 avec la détermination du périmètre fonctionnel ;

Mise en production :

- Traitements cohérence et dégroupement : mars/avril 2017 ;
- Produit Tous salariés : fin 2017.

Les charges correspondantes ont été provisionnées (non comprises dans les ressources présentées dans cette étude préalable).

4.2 Migration/prise en charge des DADS

Il est envisagé de traiter les DADS résiduelles (hors champ de l'application Siasp) dans la nouvelle chaîne.

Le scénario privilégié repose sur une intégration des données issues du Frontal-N4DS directement dans la chaîne développée pour le traitement de la DSN avec :

- accueil, contrôle et traitement de la mise au format statistique via l'application ARC ;
- traitements élémentaires : identification et codages profession et commune via le module de traitements élémentaires (à étudier : utilisation de la même capitalisation que les fichiers mensuels, ou d'une capitalisation spécifique).

Il a été vérifié que les éléments développés pour le lot 1 et en particulier le module fonctionnel d'accueil-réception-mise au format, permettra d'envisager cette solution. Toutefois, la mise en œuvre n'interviendra qu'en janvier 2017. Les adaptations et surtout les recettes seront planifiées au plus tard au second semestre 2016.

La suite des traitements sera prise en charge par le lot 2 du projet Pirénés, en particulier le traitement des salaires et des durées, ainsi que le raccord DSN-DADS (deux solutions seront étudiées : raccord direct dans le nouvel applicatif ou utilisation de la chaîne actuelle après expertise).

La maîtrise d'ouvrage a demandé la conservation de la chaine actuelle des DADS pour le traitement de la validité 2016 des DADS (les pôles DADS seront fermés en janvier 2017) comme scénario de repli pour les gestionnaires du CSSL et les traitements aval. Cette solution ne pourra être que transitoire car l'arrêt de l'application DADS est programmé pour début 2018.

Dans les deux cas, il faudra prévoir une étape de traitement de l'exhaustivité/ des doublons des déclarations en DSN ou en DADS dans le lot 2, lors de la constitution du fichier « Tous salariés » (lors de l'ajout des données en provenance des applications Siasp et PE).

4.3 Suite des projets de réingénierie

Un troisième horizon doit être géré, avec l'identification des priorités et des études statistiques à mener avant de lancer les projets futurs. Cette instruction se fait à deux niveaux :

- au niveau du comité directeur pour la priorisation des besoins des différents utilisateurs (clients) du Siera ;
- au niveau du comité opérationnel, pour la définition des projets.

Il est également nécessaire de prendre en compte le calendrier de déploiement de la DSN (par exemple sur l'entrée des employeurs publics), la qualité observée des déclarations et des données statistiques issues de la DSN pour calibrer la charge de certains développements : par exemple, sur les synthèses Employeur et Individu très dépendantes des traitements précédents et pour lesquelles les méthodes de validation sont à inventer. Un poste d'investissement a été ouvert (septembre 2015) à la division SCMT pour travailler sur l'impact du programme d'évolution sur les estimations d'emploi.

En termes de calendrier, l'ambition de la maitrise d'ouvrage est de présenter une synthèse des besoins des partenaires au comité directeur de l'automne, et une méthode de travail pour la définition des projets au comité opérationnel de fin d'année.

5 LISTE DES ANNEXES

Annexe 1 : sigles

Annexe 2 : structures de pilotage du projet

Annexe 3 : fiche de suivi du programme (version juin 2015)

Annexe 4 : compléments techniques sur la volumétrie, les performances et les tests

Annexe 1 : sigles

AA	administrateur d'application
Acoss	Agence centrale des organismes de sécurité sociale
AD	active directory
AE (déclaration sociale)	attestation employeur
APE	activité principale exercée
ARC (application)	Accueil Réception Contrôle
Artemis (application)	Application de Reprise et de Traitements Elémentaires de l'eMploI et des Salaires
AUS	architecture unifiée statistique
BDD	base de donnée
BRC (déclaration)	bordereau récapitulatif de cotisations
BRPP	base des répertoires des personnes physiques
CA	cellule Architecture
CD-PTT	comité de direction consacré à la programmation triennale des travaux
CEI	centre d'exploitation informatique
CE	Conseil d'État
Cnam	Caisse nationale d'assurance maladie
Cnav	Caisse nationale d'assurance vieillesse
Cnil	commission nationale de l'informatique et des libertés
CNIO	centre national informatique d'Orléans
COG	code officiel géographique
CSSL	centre statistique sociales et locales
DADS (déclaration ou application)	déclaration annuelle de données sociales
Dares	Direction de l'animation de la recherche, des études et des statistiques
DMMO (déclaration)	déclaration de mouvements de main d'œuvre
DR	directions régionales
Driss	département répertoires, infrastructures et statistiques structurelles
DSIJ (déclaration)	déclaration de salaires pour les indemnités journalières
DSN (déclaration)	déclaration sociale nominative
EFA	division exploitation de fichiers administratifs pour l'emploi et les revenus
EPOI	équipe de projet en organisation informatique
EPS	équipe de projet statistique
ETT	entreprise de travail temporaire
FTP	File transfert protocol
GIP-MDS	Groupement d'intérêt public modernisation des données sociales
IHM	interface homme machine
Ldap	Lightweight Directory Access Protocol
Maiol	maîtrise d'œuvre locale
MOA	maîtrise d'ouvrage
N4DS	Déclaration Dématérialisée Des Données Sociales
NIR	numéro d'identification au répertoire
ONP	opérateur national de paye

PCS (nomenclature)	nomenclature des professions et catégories socio professionnelles
PE	Particuliers Employeurs (application)
RCD	répertoire commun des déclarants
Reso	RESponsable en Organisation
RIA	responsable informatique d'application
RIA-P	responsable d'intégration des applications en production
SCMT	division synthèse et conjoncture du marché du travail
SEF	service d'échange de fichiers
SFTP	Secure File transfer Protocol
Siam	système d'identification automatique de masse
Siasp	système d'information sur les agents des services publics (application)
Sicore	système informatique de codage des réponses aux enquêtes
Siera	système d'information sur l'emploi et les revenus d'activité
Siret	système d'identification au répertoire des établissements
Sirus	système d'identification au répertoire des unités statistiques
Snip	support national Infrastructure de production
SNLA	support national des logiciels d'application
SSH	Secure Shell
UML	Unified Modeling Language
XML	eXtensible Markup Language

Annexe 2 : structures de pilotage du projet

- Note de gouvernance : n° 767/DG75-F201 du 16 avril 2014
- Séminaire de lancement de projet les 15 et 16 janvier 2015 (Compte-rendu n°56/DG75-C520 du 17 février 2015)
- Note de création des comités de pilotage et de suivi du projet Accueil et traitement de la DSN n°151/DG75-F201 et n° 25/DG75-C501 du 27 janvier 2015
- Note de périmètre de la démarche Maiol : n° 318/DG75-F201 et n° 113/DG57-C003 du 19 février 2015

Comité directeur du programme d'évolution du Siera

Réunion n°1 17/09/2014 compte-rendu n°1896/DG75-F201 du 13/10/2014
Réunion n°2 03/04/2015 compte-rendu n°711/DG75-F201 du 24/04/2015
Prochaine réunion : automne 2015

Comité opérationnel du programme d'évolution du Siera

Réunion n°1 05/12/2015 compte-rendu n°46/DG75-F201 du 12/01/2015
Prochaine réunion : 26 juin 2015

Comité de pilotage du projet Pirénés

Réunion n°1 06/03/2015 compte-rendu n°520/DG75-F201 du 25/03/2015
Réunion n°2 22/05/2015 *compte-rendu en cours*
Prochaine réunion : 18 septembre 2015

Comité de suivi du projet Pirénés

Réunion de lancement : 15/10/2014
Réunion n°1 14/04/2015 compte-rendu diffusé le 15/04/2015
Réunion n°2 11/05/2015 compte-rendu diffusé le 28/05/2015
Prochaine réunion : septembre 2015

Réunions de macroanalyse

Réunion n°1 06/01/2015
Réunion n°2 12/02/2015
Réunion n°3 30/04/2015

Réunions de travail des équipes projets

15 réunions de septembre 2014 à mai 2015

Réunions du groupe Maiol

Séminaire de lancement 10/02/2015 compte-rendu n°524/DG57-C801/AG/LB
Réunion n°2 10/03/2015 compte-rendu n°525/DG57-C801/AG/LB
Réunion n°3 24/03/2015 compte-rendu n°527/DG57-C801/AG/LB
Réunion n°4 14/04/2015 compte-rendu n°528/DG57-C801/AG/LB
Réunion n°5 21/05/2015 *compte-rendu non diffusé*
Prochaine réunion : 18/06/2015

Réunions de travail avec le CEI

Réunion n°1 06/01/2015
Réunion n°2 24/03/2015
Réunion n°3 30/04/2015
Prochaine réunion 25/06/2015

Annexe 3 : fiche de suivi du programme d'évolution du siera

Dsds	**Programme d'évolution du Siera** **Projets de réingénierie**	SIERA

Date de mise à jour de la fiche : juin 2015

1. Suivi du projet externe DSN

Production : avec l'application du décret d'obligation intermédiaire, 75% des entreprises concernées ont déposé une DSN en mai 2015 (5 millions de salariés couvert), les trois quarts des dépôts se font en DSN phase 1, un quart en DSN phase 2.

Phase 1 (objectif : substitution des DMMO, DSIJ, AE[28]) : la montée en charge s'est accélérée, à partir de mars, avec 3,7 millions de salariés couverts sur la déclaration d'avril (déposée en mai 2015).

Phase 2 (objectif : remplacement du BRC Acoss et intégration des entreprises de travail temporaire-ETT)

Bilan du pilote (prévu de novembre 2014 à février 2015, prolongation d'un mois) :

- une trentaine d'entreprises sur 70 prévues dans le pilote, ont déposé une DSN avec succès ;
- aucune « grosse » entreprise n'a participé au pilote;
- une seule ETT est parvenue à déclarer mais qualité insuffisante ;
- des alertes sur le RCD (référentiel des cotisants/déclarants) avec des rejets à tort d'établissements.

Qualité des données inégale :

- 50% de déclarations de bonne voire, d'excellente qualité, 50% de qualité médiocre, voire mauvaise sur les données du pilote de la phase 2 (du point de vue du recouvrement) ;
- toujours des difficultés sur l'utilisation des blocs changements (Dares, Pôle emploi).

Décision de mise en production à compter du 17 mars de la phase 2 (CD DSN du 6/03/2015) mais sur un champ plus restreint (report pour les ETT).

Prochaines échéances

Phase 1 : disparition de la DSN - phase 1, obligation de basculer en phase 2 à compter de septembre 2015

Phase 3 : pas de modification du calendrier

- octobre 2015 -janvier 2015 : mise en place d'un pilote
- février 2016 : mise en production de la phase 3

Textes de référence sur la DSN

- *Article 35 de la Loi « Warsmann » n°2012-387 du 22 mars 2012 relative à la simplification du droit et à l'allègement des démarches administratives (JORF du 23 mars 2012)*
- *Décret 2013-266 du 31 mars 2013 relatif à la déclaration sociale nominative (phase 1)*
- *Décret n° 2014-1371 du 17 novembre 2014 relatif à la déclaration sociale nominative (phase 2)*
- *Décret n° 2014-1082 du 24 septembre 2014 fixant les seuils de l'obligation anticipée d'effectuer la déclaration sociale nominative*

[28] DMMO : déclaration de main d'œuvre (Dares), DSIJ : déclaration de salaires pour les indemnités journalières (Cnam), AE : attestation d'emploi pour Pôle emploi.

2. Travaux entre l'Insee et le projet DSN

Avril 2015 : validation du **cahier des charges** pour la phase 3 (variables demandées, contenu des flux, « tuyaux » entre la Cnav et l'Insee).

Financement DADS/DSN : les principes de tarification retenus pour la DSN devraient être proches de ceux existants pour les DADS. Le principe d'un gel des montants 2014 a été proposé pour couvrir la période de concomitance DSN/DADS. Une fois le système entièrement couvert par la DSN, la participation de l'Insee pourrait être revue à la baisse suite à l'augmentation du nombre de partenaires (réunion TDS du 28 mai 2015, en attente du compte-rendu).

Prochaines échéances

- Accès aux données DSN-phase 2 pour fichiers tests en juillet 2015.

Points de vigilance pour l'Insee (vus avec la MOAS lors de la réunion du 13/05/2015)

- travaux avec la direction de la Sécurité sociale (décret phase 3 et arrêté filtre) et dossier Cnil pour la phase 3 ;
- raccord au pilote de la phase 3 (demande effectuée).

3. Extension du champ et coproduction des ETE

Avril- juin 2015 : réunions Acoss-Dera sur :

- les **flux** à mettre en place (cadre co-production et montée en charge de la DSN) ;
- les règles opérationnelles de **partage du champ** de production ;
- le choix des méthodes de cvs, avec la participation du DMS.

Au second semestre démarreront des réunions sur la dimension des échanges **Urssaf-directions régionales Insee** à mettre en place : première réunion de cadrage du GT (à lancer en septembre) programmée fin juin (Acoss-DAR-Dera).

Convention Insee-Acoss prévue fin 2015.

4. Avancement du programme d'évolution du Siera

27/05/2015 : avis du comité des investissements sur l'expression des besoins

Juin 2015 : remise de l'étude préalable du projet Pirénés - lot 1 (ex- Accueil et traitement de la DSN)

Agenda

26/06/2015 : 2ème comité opérationnel du programme d'évolution du Siera

18/09/2015 : 3ème comité de pilotage du projet Pirénés

Automne 2015 : 3ème réunion du comité directeur du programme d'évolution du Siera

Démarche Maiol

Rapport sur l'analyse de l'existant diffusé en juin

Prochaine échéance : tests utilisateurs sur l'IHM gestionnaire (juillet).

Avancement du projet Pirénées (ex-Accueil et traitement de la DSN)

Conceptions générales terminées ou en cours de finalisation (cf. compte-rendu du 2ème comité de pilotage à venir).

Application ARC (accueil, réception, contrôle, mise au format des données, gestion des nomenclatures)

Version 2 en cours de recette. Tests de performance à venir. L'application est quasiment achevée.

Application ARTEMIS (traitements statistiques élémentaires, gestion des référentiels)

Spécifications et développements des automates de codage (y compris appels aux services externes) en cours.
Développement des premières IHM des gestionnaires.

Annexe 4 : compléments techniques

La volumétrie

Les estimations de la volumétrie des espaces de réception et de stockage des données - en production - qui sont présentées ici concernent l'architecture cible à horizon 2020 (i.e. à compter du traitement de la validité 2020).

- On peut, en effet, distinguer trois périodes de montée en charge de la DSN selon les dates d'entrée des entreprises dans le système:
  - à court terme (CT) , validités 2016 et 2017, sont concernées par la DSN les entreprises du privé hors titres simplifiés,
  - à moyen terme (MT), validités 2018 à 2019, s'y ajoutent les entreprises du public hors titres simplifiés,
  - à long terme (LT) correspondant à la cible, validités 2020 et suivantes, tout le champ salarié hors particuliers employeurs.

Type de déclarations	CT	MT	LT
Mensuelles (DSN)	415 000	493 700	1 897 000
Annuelles (DADS)	1 335 000	1 403 300	

Évolution du nombre de déclarations reçues par type au court de la période 2016-2021

Les estimations de volumétrie **des espaces de stockage des fichiers** sont calculées en prenant en compte les éléments suivants :

- La Cnav livre chaque mois 1 fichier DSN par établissement - ou déclaration - (les discussions sont en cours avec le GIP-MDS pour préciser le contenu, la taille et les conditions d'envoi). On estime à 1 897 000 le nombre d'établissements dans la configuration cible (validité 2021 et suivantes).

- 17 058 fichiers de test phase 1 reçus en octobre 2014 occupent sur serveur un espace de 0,13 Go zippés (pour 1,25Go non zippés soit un facteur 10). C'est donc 23 Go qui seront livrés mensuellement soit annuellement, à la cible, à partir de la validité 2021, 12*23 Go soit **273 Go**

Base de données

Les estimations de volumétrie de la base de données sont calculées en prenant en compte les éléments suivants :

ARC

- 17 058 fichiers test chargés en base occupent un volume de 1,7 Go. Les 1 897 000 déclarations mensuelles chargées devraient occuper mensuellement un volume de (1 897 000/17 058)*1,7 Go = 190 Go.
- Les données reçues sont dupliquées 6 fois lors des 5 phases de chargement (chargement, contrôle, normage, filtrage, mise au format) pour être visible à toutes les étapes du processus. Le volume mensuel nécessaire au stockage en base des données brutes s'élève donc à 190 Go*5 ce qui représente environ un 1,15 To par mois.
- En régime courant au plus 3 mois de données brutes stockées en base de données ; les données traitées sont détruites au fur et à mesure.
- **Le volume de données stockées dans la base de données de ARC s'élèvera donc à approximativement 3,5 To en 2020.**

Artemis

- Les données utilisées par Artemis représente au plus 40% des données brutes : seule une petite partie des variables est retenue. L'application Artemis retiendra 25 mois de données soit un volume total de 190 Go * 25 * 0.4= 1,9 To en base de données.

- Pendant les phases de court et moyen terme, le chiffrage tient compte du reliquat de DADS à traiter. Sous l'hypothèse que le volume d'une DADS est à peu près équivalent au volume d'une DSN mensuelle, le volume des données chargées annuellement en base est donc estimé à **486 Go** à CT, **564 Go** à MT.
- Artemis doit aussi créer différent produits représentant environ 500Go de données
- Artemis va aussi héberger des référentiels de capitalisation dont le volume ne dépassera pas 50 Go de données
- La base de l'IHM s'élèvera aussi à environ 50 Go de données
- **Le volume de données stockées dans Artemis à terme s'élèvera donc à approximativement 3 To**

- Volumétrie liée aux index : les éléments de calcul suivants sont pris en compte :

  - Les spécificités de chacune des applications, ARC ou Artemis, sont également prises en compte :
    - ARC traite en batch des données brutes et nécessite pas ou peu d'index (quelques clés primaires)
    - Seule la table de pilotage nécessite la pose d'index spécifiques pour optimiser sa lecture via le poste de pilotage : **ratio estimé de 1,001**
  - Dans le cas d'Artemis, les besoins d'indexation seront différents selon que les données seront traitées par le batch ou l'IHM mais un **ratio moyen de 1,05** semble raisonnable.

Ces éléments sont résumés dans le tableau suivant :

Module	Volume des données	Ratio appliqué	Volume des index
ARC	3,45 To	0.001%	4 Go
Artemis	3To	7%	200 Go

Espace de stockage des fichiers utilisés par les applications

Cet espace contient :

- les fichiers DSN fournis par la Cnav reçus depuis 4 mois : 23 Go x 4 mois = 100 Go
- les fichiers des autres sources de données à charger (DADS,…) = 100 Go

L'espace de stockage est défini sur un serveur de fichier accessible par ARC. Les utilisateurs peuvent récupérer les fichiers via l'IHM de l'application ARC utilisée sur l'espace AUS.

Espace de conservation de fichiers

Cet espace contient les fichiers ou les produits réalisés pendant 5 ans et qui ne sont plus utiles à la chaine applicative. Les utilisateurs peuvent en revanche y accéder en lecture seule.

En détail, cet espace contient :

- les fichiers de déclarations livrés mensuellement par la CNAV depuis 5 ans - 3 mois : 1,5 To
- les fichiers de déclarations chargées et normées dans ARC depuis 5 ans : 1,5 To
- les autres sources de données brutes (DADS, …) : 0,5 To
- les autres sources de données chargés et normées dans ARC (DADS, …) : 0,5 To
- les produits d'Artemis sur 5 ans : 1 To

Clone

- La MOA souhaite la mise à disposition d'une année de données (brutes ou statistiques) pour analyse. Pour les données brutes, l'interface de ARC devrait permettre de s'affranchir de l'utilisation d'un clone. En revanche, la base de données Artemis devrait vraisemblablement être clonée soit 3,5 To de données.

Autres espaces

- Les espaces nécessaires au développement, tests fonctionnels et recette sont estimés globalement à 1To (ces environnements sont déjà créés au CEI).
- Une plateforme sera mise à disposition pour réaliser des tests de performance offrant l'espace nécessaire pour stocker les fichiers de test et une base de données ayant les caractéristiques de le base de données de production

➢ Ces éléments sont résumés dans le tableau qui suit (pour la période 2016 - 2020) :

Espaces de stockage	*Volume réel estimé*
Production	
- Espace de stockage applicatif partagé serveur Batch	
*Stockage et gestion réception et conservation des fichiers livrés par la CNAV durant la durée de rétention de ARC (2*10^6 fichiers/mois sur 3 mois)*	*0,1To*
Autres sources de données (DADS, …)	*0,4To*
Total espace de stockage	***0,5To***
- Espace de conservation de fichiers	
*Conservation des fichiers livrés par la CNAV (2*10^6 fichiers/mois sur 5 ans - 3 mois)*	*1,5To*
*Conservation des fichiers chargés dans ARC (2*10^6 fichiers/mois sur 5 ans - 3 mois)*	*1,5To*
Conservation des autres sources de données brutes (DADS, …) sur 5 ans - 3 mois)	*0,5To*
Conservation des autres sources chargées dans ARC (DADS, …) sur 5 ans - 3 mois)	*0,5To*
Conservation des fichiers produits par l'application Artemis sur 5 ans	*1 To*
Total espace de conservation	***5 To***
- base de données	
base de données ARC	*3,5 To*
base de données Artemis	*3,5 To*
Total base de données Production	***7 To***
Clone base de données pour expertises MOA (éventuel)	*3,5 To*
Total clone	***3,5 To***
Total Production	***16 To***
Autres environnements(développement, tests fonctionnels, recette Intégration)	
- base de données de développement, tests fonctionnels, recette, intégration	*1 To*
- base de données tests de performances(batch) ponctuellement et espace de stockage	*3,5 To*
Total Autres plateformes	***4,5 To***

Les performances : propositions et solutions envisagées

Les exigences de performance sont dictées par un calendrier de production mensuel très contraint au vu des volumes de données à traiter mensuellement (campagne mensuelle).

Réduire les volumes

Un objectif : obtenir de bonnes performances malgré un volume conséquent de données à traiter (3,5 To par an). Les solutions envisagées ou déjà mises en œuvre sont à 2 niveaux :

a) Les solutions "métier"

- Limiter dans la base de données le nombre de validités en cours de gestion
  - Artemis : nombre limité à 3 validités mensuelles

- Réduire à 5 ans sur les espaces de stockage les années de conservation des données

- Limiter la redondance des données
  - ARC : seules les données brutes et leur version normée sont conservées dans la base de données. Les données produites durant les autres étapes du processus ne sont pas conservées.
  - Mutualiser les services applicatifs de répertoires et de nomenclatures entre ARC et les applications clientes (Artemis).

- Limiter le nombre de variables retenues
  - Le nombre de variables métier est bien inférieur au nombre de variables brutes par utilisation de filtres et agrégations (rémunérations, durées des contrats, etc)

- Nombre de variables Artemis limité
  - Par exemple, seules les variables déclarées identifiantes des variables à coder sont retenues par exemple à un code employeur sera identifié par le code siret déclaré, la raison sociale et le depcom

b) Les solutions techniques

- Compression des données
  - Compresser systématiquement les données à télécharger (ARC)
  - Traiter directement les données sans décompression (chargement ARC)

Améliorer les performances

a) solutions techniques

- Utiliser les solutions connues et éprouvées mais adaptées au nouveau système de base de données PostgreSQL
  - Partitionnement par famille de normes, sur nom des fichiers (à étudier)
  - Indexation ciblée. L'indexation sera par ex. plus importante sur les données alimentant le poste du gestionnaire que sur celles utilisées par les traitements batch.

- Implémenter les solutions préconisées et adoptées par ailleurs (Harmonica)
  - Typage des variables (option Collate)

- Initialisation de contextes d'exécution limités aux seules données utiles au traitement

- Parallélisations des traitements

  - Une version fonctionnelle existe dans les deux versions : traitements séquentiels ou parallélisés, confronter aux tests de performances.

Les questions qui restent posées

1. Dans un premier temps, le SNIP et la cellule architecture ont validé la solution de monter une base de données de 4To sans partitionnement sur laquelle pourront être réalisés les tests de performance. Le CEI va réaliser cette opération.
   Le SNLA étudie la possibilité de mettre en place et gérer des datastores sur des pointeurs compressés.

   La contrainte n'existe pas pour les fichiers (les têtes NAS n'ont pas de limitation de taille).

   Par ailleurs, les opérations de sauvegarde et de maintenance quotidiennes réalisées par le service d'exploitation ne sont pas à négliger (impact des volumes de la base de données sur les temps de sauvegarde, conséquences sur la durée des calculs de statistiques etc.)

2. Les tests de performances permettront de valider les solutions qui seront in fine retenues et pour lesquelles un choix définitif n'est pas fait : le partitionnement ou la parallélisation des traitements.

Tests unitaires et tests fonctionnels

Outre ceux décrits au § 1.7.4, il est prévu la réalisation des tests suivants dans le cadre du projet Pirénés :

Les tests unitaires

À la charge du développeur, ils permettent de tester à travers des classes de test spécifiquement dédiées à cela les classes de service (traitement Java sur des variables Java) et les classes DAO (accès et traitement des données stockées en BDD). Construit à partir des Frameworks préconisés par la cellule architecture JUnit et DBUnit.

L'objectif sera de couvrir par les tests unitaires 20% du code et d'utiliser le déploiement de l'application sur la PIC pour vérifier à fréquence régulière la qualité des programmes et la non-régression.

Les tests fonctionnels

A la fois utilisés par l'EPOI au moment de l'intégration fonctionnelle et par l'EPS pour réaliser les tests de recette, ils sont un gage de la qualité applicative.
Le choix a été fait de préparer ces tests conjointement entre l'EPOI et l'EPS c'est à dire partager une méthodologie, préparer le plan de test et les données de test en entrée et de résultats attendus.
Cependant, chaque acteur se réserve la possibilité de compléter avec ses propres jeux de tests. Ils seront nécessairement documentés ainsi que le résultat du passage des différentes campagnes de test.
Il est à noter que le poste de pilotage et les batchs feront l'objet de tests fonctionnels

L'EPOI bénéficiera de la plateforme de la ressource test du CNIO spécialement dédiée aux tests fonctionnels.
L'EPS utilisera une plateforme de Qualification Fonctionnelle du projet au CEI.