Culture
ARCHIVES NATIONALES

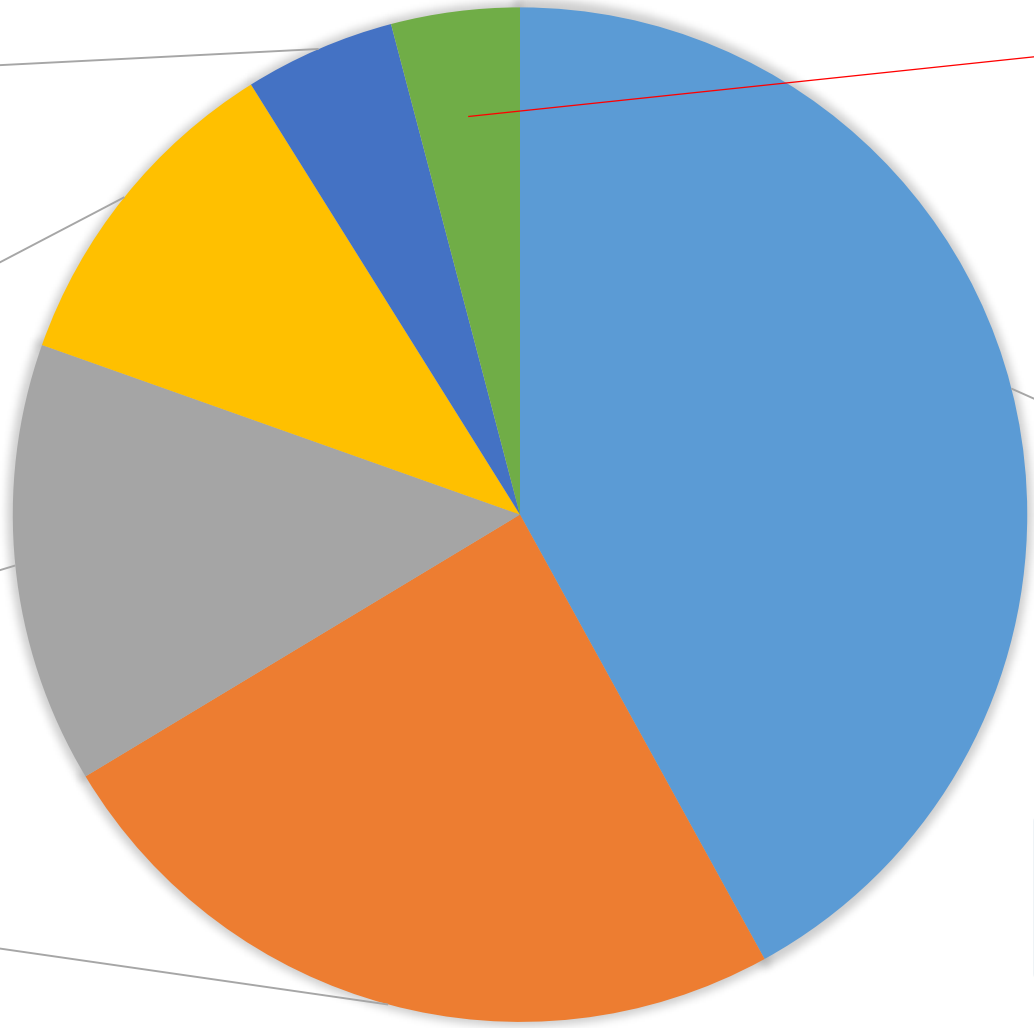
2 décembre 2021

La reprise du patrimoine numérique des Archives nationales (1983-2018) : les données issues des enquêtes statistiques et leurs métadonnées

La reprise des données de Constance vers le SIA : les principes

- Une histoire qui s'inscrit dans la continuité des 6 migrations précédentes
- Garder les données et les métadonnées intègres, authentiques, lisibles et accessibles dans le temps
- Rassembler toutes les données et les métadonnées pour créer les moyens de les faire converger vers le SIA numérique

Le patrimoine numérique des AN (1983-2018)



Archives audiovisuelles (27)

Archives photographiques (23)

Messageries (60)

Enquêtes statistiques (236)

Extractions de bases de données (79)

Archives bureautiques (137)

Total : 562 versements (73 To)

Nota : ces chiffres comprennent les versements déjà repris, les versements en cours de reprises « industrialisées », et ceux hors périmètres des reprises « industrialisées »

Organisation du chantier de reprise de données

1. Définir des lots présentant des caractéristiques communes permettant d'industrialiser le plus possible les processus de reprise
2. Les premiers chantiers de la reprise de données concernent donc les archives qu'il était possible de **reprendre de manière la plus industrielle possible**.
   - En reprenant des données archivées de manière homogène, nous pouvons penser **la fabrication en masse de SIP** eux aussi homogènes.
   - Pour certains types de données, il est même possible de constituer des **profils d'archivage** pour aider à la fabrication et au contrôle des SIP au moment de leur transfert dans le SIA numérique
3. Reprendre des versements plus spécifiques, nécessitant des traitements scientifiques et techniques particuliers.

Historique de l'archivage numérique aux Archives nationales

Les enquêtes statistiques, premières archives numériques

Première collecte d'archives numériques en 1983, après la mise en place de **CONSTANCE**

CONSTANCE (CONServation et Traitement des Archives Nouvelles Constituées par l'Électronique) = politique, processus et **méthode de traitement et de conservation des données numériques et de leurs métadonnées**

Collecte des données des enquêtes qui ont constitué **les premières bases informatiques** créées par des ministères et des institutions pionniers dans ce domaine comme **l'INSEE** et **l'INED**, ou encore **le ministère de l'Agriculture** et l'ancien **ministère des Transports**.

Les enquêtes reflètent les politiques de l'Etat pour analyser, étudier, quantifier le monde contemporain et ainsi soutenir et justifier des choix stratégiques. L'étude de ces versements des grandes enquêtes de la France permet également de saisir l'évolution de la politique économique, industrielle, agricole et sociale.

Historique de l'archivage numérique aux Archives nationales
La méthode Constance

1) Pérenniser les fichiers de données sur des bandes LTO (stockage à froid), à plat, et nommés de manière homogène :
NoticeProducteur_NuméroVersement_NumeroArticle_NomFichier.ExtensionMaison

Les **données** :

- Représentent les réponses aux questionnaires
- Encodées en ASCII (American Standard Code pour Information Interchange, norme d'encodage des caractères) selon le principe 1 donnée = 1 à n caractères, et il y a autant d'octets que de caractères
- Archivées à plat dans des documents de type texte et stockées sur bande LTO
- Les fichiers de données sont gérés et décrits dans une base documentaire

Historique de l'archivage numérique aux Archives nationales

La méthode Constance

2) Les **métadonnées** (informations de représentation) sont saisies dans des bases documentaires :

Elles sont réalisées grâce à une association entre le producteur et les Archives Nationales et sont indispensables pour accéder aux données et les comprendre.

Dans les bases de données CINDOC ont été conservées les fiches d'application (information sur l'application versée), dictionnaires de données, fiches de structure, dictionnaires de codes

3) La **documentation associée papier** (numérisée) est conservée :

Elle contextualise la production des données (instructions aux enquêteurs, questionnaires d'enquête vierges, bilan de l'enquête, publications de la recherche) et documente les échanges entre le producteur et les AN

Cet ensemble fournit la signification des données et décrit comment les données sont représentées dans les fichiers.

Un exemple
Enquête mobilité géographique et insertion sociale en 1992 (AN, 19990409), enquête de l'INSEE : une enquête sur la vie des immigrés et de leurs enfants en France.

Les données

```
FR1101702120069711911350002 090 01 01131129365012 1              1
00110170200697205110UI911350002
11000115336170007732070001102202031030112360123531         1   12                          12
0111017022            00               12222  4           0 01    992261750560122021122    0122   22222299012012012012220698788883309080100000004111   2  462
250  25025014115211984690113120172250101721220311017020100202
041101702010272KARIM          12509811
051101702011        320031  270852  1 11    2 2
061101702KABYLE                   170FRANCAIS
11122321211155121221221223322
2                           2         1                    2623503222062511121      1220203900500773207107010170201037072682843392112607110170202036870682843392112607110170203035767672553212112610110170204015357120526740
```

Les métadonnées

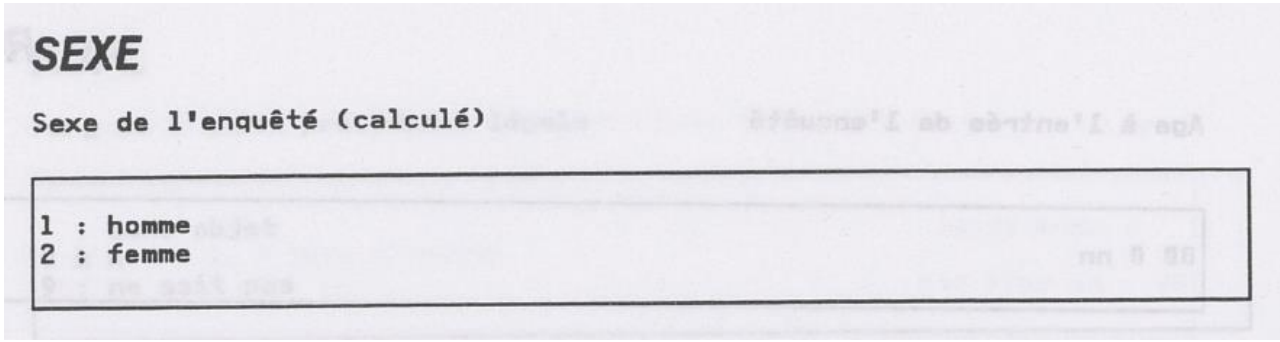
Extrait de la fiche de structure

SYMB	N-STRUCT	POS-DEPA	LONG	POS-ARRI	TYP-DONN
TYPE	Enregistrement 00	1	2	2	
IDENT		3	7	9	A.N.
NENQ		10	5	14	A.N.
SEXE		15	1	15	A.N.
JOUR		16	2	17	A.N.
MOIS		18	2	19	A.N.
CONF		20	3	22	A.N.
D		23	2	24	A.N.
C		25	3	27	A.N.
CIL		28	3	30	A.N.
IL		31	4	34	A.N.
FIL		35	1	35	A.N.
NLOT		36	3	38	A.N.
NCHIF		39	4	42	A.N.
R1		43	1	43	A.N.
R2		44	2	45	A.N.
R3		46	2	47	A.N.
R4		48	2	49	A.N.
R5		50	2	51	A.N.
R6M		52	2	53	A.N.
R6A		54	2	55	A.N.

Extrait du dictionnaire des données

NOM	A-DEB	A-FIN	SYMB
SEXE DE L'ENQUETE (CALCULE)	1992	1992	SEXE

Extrait des codes



ARCHIVES NATIONALES

Un exemple
Enquête mobilité géographique et insertion sociale en 1992 (AN, 19990409), enquête de l'INSEE : une enquête sur la vie des immigrés et de leurs enfants en France.

```
FR1101702120069711911351   0002 090 01 01131129365012 1              1
00110170200697205110UI91135   0002
11000115336170007732070001102202031O3
                              0112360123531         1   12                          12
0111017022            00                12222  4           0 01    992261750560122O2
11              0122   22222299012012012012222                    6987888833090801000000041111   2  462
250  250250141152119846901l 3120172250101721 22
031101702    0100202
041101702010272KARIM          12509811
051101702011        320031  270852  1 11    2 2
061101702KABYLE                   170FRANCAIS
111223212111551212212212233 22
2                             2        1                   2623503222062511121      1220203900500773 2
071101702010370726828433921126
071101702020368706828433921126
071101702030357676725532121126
1011017020401535 71  20526740
```

Le sexe de l'enquêté se trouve en position 15 (15e caractère de l'enregistrement), et est de longueur 1.
1 = homme, 2 = femme
Dans cet enregistrement, la personne enquêtée est une femme.

La reprise automatisée des données d'enquêtes statistiques
Travaux préparatoires

Rendre les **données accessibles** en effectuant (été 2019) :

- Le transfert depuis les bandes LTO vers le nouveau serveur installé sur le site de Pierrefitte-sur-Seine ;
- La sortie des données de leur encapsulage (pour d'autres typologies de données)

Rendre les **métadonnées accessibles** en les réunissant dans une **cartographie** (2017-2019) :

- Métadonnées extraites de CINDOC
- Métadonnées provenant du SIA
- Intégration des informations issues des nouveaux référentiels (règles de communicabilité, par exemple)

Numériser la documentation associée

La reprise automatisée des données d'enquêtes statistiques

Objectif

L'existant :

- Données archivées à plat, avec un nommage homogène des fichiers
- Métadonnées de description et de représentation extraites de CINDOC, métadonnées de gestion dans le SIA, …, toutes rassemblées dans une cartographie

La cible :

- Constituer des paquets d'information à verser (dans l'OAIS « SIP ») conformes au **SEDA** (Standard d'échanges de données pour l'archivage)
- Verser ces paquets dans les modules du SIA numérique.

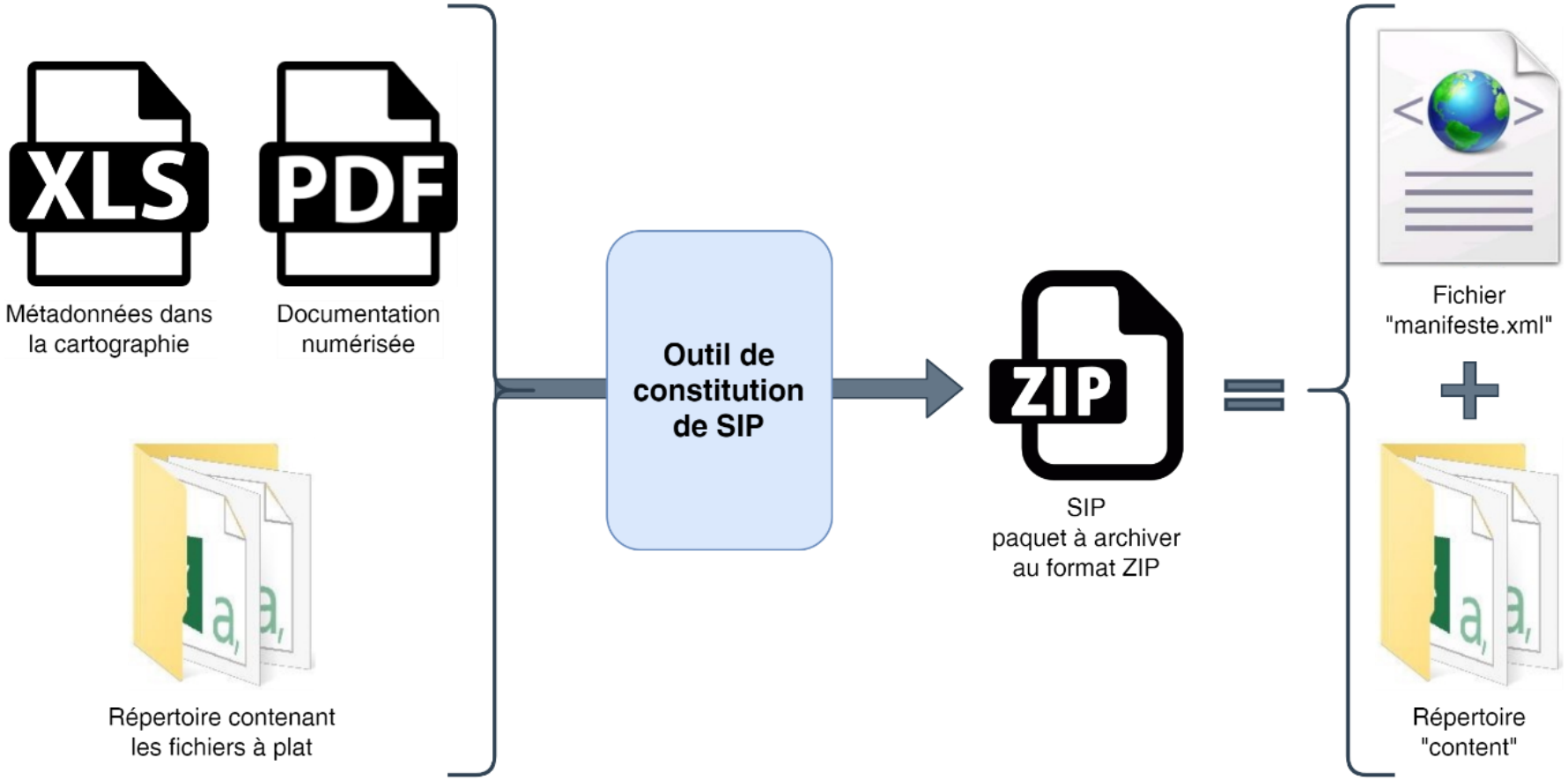
La reprise automatisée des données d'enquêtes statistiques
Périmètre

Les données d'enquêtes des 4 producteurs historiques d'archives électroniques (INSEE, INED, ministère de l'Agriculture, ministère des Transports)

150 versements pour 134 applications, un peu plus de 5 000 fichiers de données soit 150 Go, 2 065 fichiers correspondant à la documentation associée en PDF soit 116 Go (travail non fini)

- Les **données** des enquêtes statistiques
- Les **informations de représentation** (INDISPENSABLES !)
- La **version numérisée de la documentation associée** aux entrées de données statistiques permettant de comprendre le contexte de production et de documenter les opérations d'archivage (ex : questionnaire d'enquêtes vierges, instructions aux enquêteurs, résultats de l'enquête, etc.)
- Les **métadonnées de description et de gestion** de ces enquêtes saisies entre 1983 et 2017 dans les bases Cindoc et transférées depuis dans la cartographie.

Le bordereau de transfert (dit "manifeste"), fichier XML conforme au SEDA 2.1, liste les unités archivistiques, avec les métadonnées descriptives, techniques et de gestion associées, et les fichiers à archiver (ainsi que leurs identifiants uniques), et permet d'en contrôler le nombre.



Le répertoire "content" contient tous les fichiers à archiver, stockés à plat (sans arborescence), et renommés avec des identifiants uniques, produits par l'outil de constitution de SIP.

La reprise automatisée des données d'enquêtes statistiques
Modélisation des paquets à constituer

AMOA avec le cabinet Mintika

1) **Modélisation du paquet « type »** (= la structure du paquet + les métadonnées associées à chaque objet archivé)
2) **Élaboration du profil d'archivage (XML)**

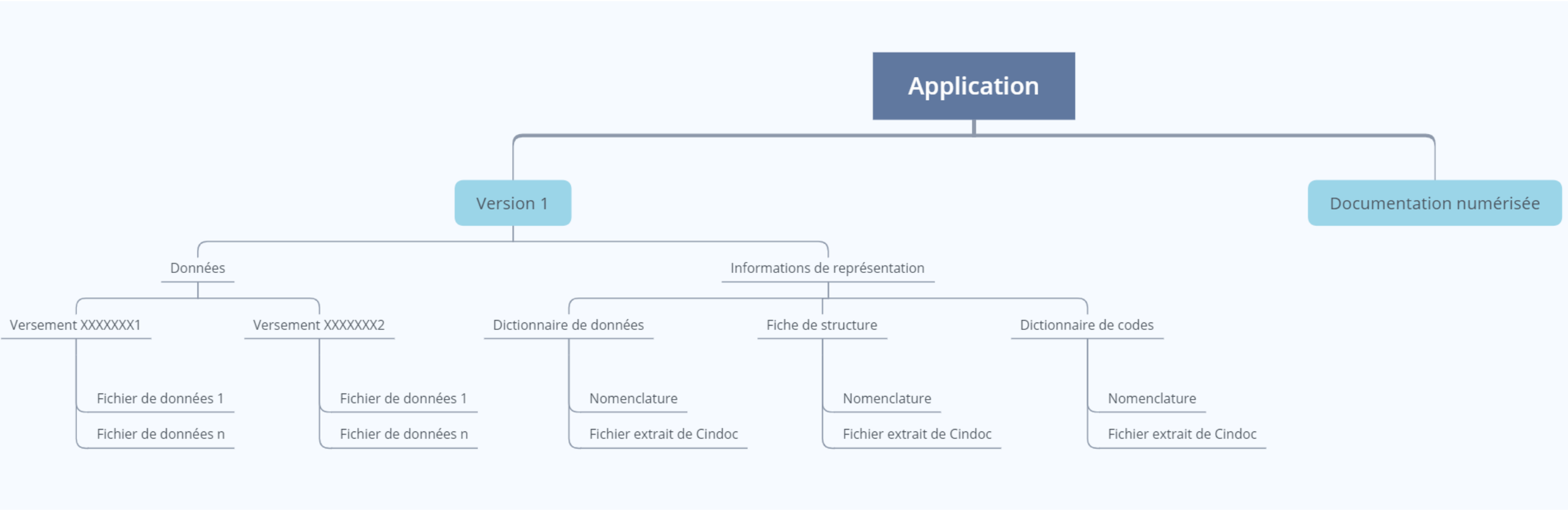
Solution qui **regroupe des données et des métadonnées d'une même enquête (« application » dans Constance) dans un paquet.**

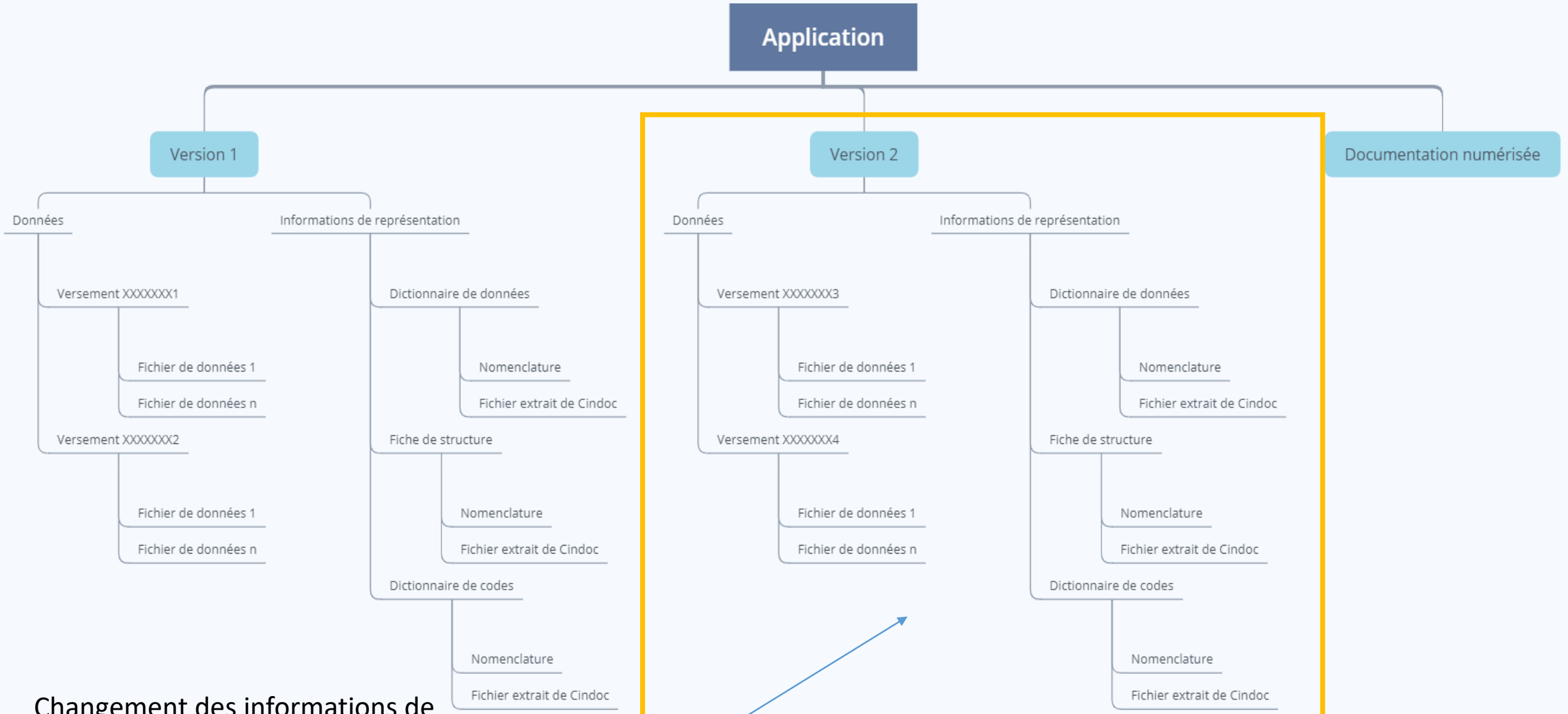
1 SIP = 1 application/enquête qui a pu faire l'objet de plusieurs opérations d'archivage => regroupement des données archivées à différents moments (plusieurs versements)

Meilleure lisibilité et exploitabilité des données par rapport à leurs informations de représentation

La reprise automatisée des données d'enquêtes statistiques

Modélisation des paquets à constituer





Changement des informations de représentation pour s'adapter à l'évolution des données ou pour insérer une nouvelle entrée pour une même enquête

La reprise automatisée des données d'enquêtes statistiques
Développements et construction des SIP

Deux développements ont servi à la création des fichiers prévus par le modèle de données pour assurer le transfert des données vers le SIA numérique (fichiers d'informations de représentation et PDF de documentation)

- Un 1er programme convertit les ensembles d'images de documentation associée au format JPEG en fichiers PDF, plus faciles à consulter.
- Le 2ème programme crée des fichiers au format CSV contenant les informations de représentation s'appliquant à une enquête.

Le 3ème programme consiste, pour chaque enquête, à générer de manière automatisée un SIP **à partir des fichiers concernant une enquête et des métadonnées s'y référant dans la cartographie**.

Exemple d'une enquête transférée en recette

La reprise automatisée des données d'enquêtes statistiques

Transfert en production depuis septembre 2021 (86 enquêtes versées)

Suite des transferts en 2022, après retour du dernier lot de numérisation

Reprise des enquêtes réalisées par d'autres ministères

Reprise d'un deuxième lot « industrialisé » : les données décrites par des instruments de recherche EAD

ARCHIVES NATIONALES

Merci de votre attention !
Des questions ?

Emeline Levasseur : emeline.levasseur@culture.gouv.fr