Direction interministérielle
des systèmes d'information
et de communication

Cadre Commun d'Architecture des Référentiels de données

Complément n°2
au Cadre Commun d'Urbanisation
du Système d'Information de l'État version 1.0

Version n°1.0 du 18/12/2013

Table des matières

1. Synthèse 4
2. Objectifs du cadre commun d'architecture 6
2.1. Pourquoi un cadre commun d'architecture des référentiels de données ? 6
2.1.1 Un axe stratégique de transformation du SI de l'État 6
2.1.2 Constats partagés 7
2.1.3 Objectifs 8
2.2. Portée du cadre d'architecture 9
2.3. Articulation avec le corpus réglementaire 9
2.4. Entretien et mise à jour du document 10
2.5. Version du document et références documentaires 10
2.6. Diffusion 11
2.7. Approbation 12
3. Définitions et principes 13
3.1. Une donnée 13
3.2. Sémantique et modèle sémantique 13
3.2.1 Définir pour comprendre et se comprendre 13
3.2.2 Monter en abstraction pour simplifier, rendre plus robuste et flexible 14
3.2.3 Standardiser pour l'interopérabilité 15
3.2.4 Formaliser et modéliser 15
3.3. Les données de référence 16
3.4. Une métadonnée 17
3.5. Qualité des données 17
3.6. Un référentiel de données 18
3.7. Urbanisation du SI de l'État et référentiel de données 21
4. Règles d'architecture de données de référence 23
4.1. Règles de niveau Stratégique 23
4.2. Règles de niveau Métier 26
4.3. Règles de niveau Fonctionnel 27
4.4. Règles de niveau Application 29
4.5. Règles de niveau Application et Infrastructure 33
5. Architectures types d'un référentiel de données 34
5.1. Architecture type d'un référentiel 34
5.2. Pattern 1 : référentiel centralisé 35
5.3. Pattern 2 : Référentiel de consolidation 36
5.4. Pattern 3 : Référentiel de coopération 37
5.5. Pattern 4 : Référentiel esclave 38
5.6. Pattern 5 : Référentiel hub 38
6. Mise en place & Gouvernance 40
6.1. Les étapes de mise en place d'un référentiel 40
6.2. Les activités de gouvernance, de gestion et de MCO d'un référentiel 42
A Check-list de la conformité au cadre 44
B Modèle de catalogue de référentiel et de services associés 45
C Glossaire de la gouvernance des données 47

Liste des figures

Figure 1 - Avantages et objectifs d'une gouvernance des données *8*
Figure 2 - Articulation du Cadre Commun d'Architecture des Référentiels de données avec l'ensemble du corpus réglementaire *10*
Figure 3 - Tableau des contributeurs à la rédaction du Cadre Commun d'Urbanisation du SI de l'État *11*
Figure 4 - Qualité des données : les critères intrinsèques *18*
Figure 5 - Qualité des données : les critères de services *18*
Figure 6 - Qualité des données : les critères de sécurité *18*
Figure 7 - Périmètre, couverture et portée d'un référentiel de données *19*
Figure 8 - Architecture générale d'un référentiel de données *19*
Figure 9 - Acquisition, point de vérité et consommation des données *20*
Figure 10 - Les six axes majeurs d'une gouvernance des données *22*
Figure 11 - Exemple de dépendances entre données de référence de la sphère sociale *28*
Figure 12 - Synchronisation de données de référence : une pratique à encadrer *31*
Figure 13 - La recopie d'une copie de données de référence est interdite *31*
Figure 14 - Architecture logique type d'un référentiel *34*
Figure 15 - Pattern 1 : Référentiel centralisé *35*
Figure 16 - Pattern 2 : Référentiel de consolidation *36*
Figure 17 - Pattern 3 : Référentiel de coopération *37*
Figure 18 - Pattern 4 : Référentiel esclave *38*
Figure 19 - Pattern 5 : Référentiel hub *39*
Figure 20 - Le cycle général et les étapes de mise en place d'un référentiel *40*
Figure 21 - Les activités de gouvernance et de gestion d'un référentiel *42*
Figure 22 - Check-list de conformité au présent cadre *44*
Figure 23 - Structure de données d'une Application de type référentiel *45*
Figure 24 - Structure de données d'un Objet métier *46*
Figure 25 - Structure de données d'un Attribut d'un objet métier *46*
Figure 26 - Structure de données d'un Service *46*

1. SYNTHÈSE

L'enregistrement et le traitement de la donnée sont la manifestation d'une transaction (paiement, déclaration de revenus, …), d'une réalité physique (information topographique par exemple), d'un objet légal (cadastre, immatriculation des entreprises, …), d'une relation de service (demande d'information, inscription, …), d'une décision (autorisation, gestion de droits et d'habilitation, …). L'automatisation de plus en plus systématique des processus donne un rôle prépondérant à la donnée numérique et dans un nombre croissant de cas, **la donnée numérique est devenue la seule manifestation et la seule trace** de la transaction, de la décision, de l'objet légal. **Les données constituent un des principaux actifs stratégiques de l'État**.

Considérer la donnée comme un actif nous invite à considérer les questions suivantes : Comment optimiser son coût d'acquisition ? Comment entretenir cet actif et créer le maximum de valeur avec cet actif ? Comment le partager efficacement ?

Le recueil de données présente **un coût d'acquisition pour l'État**, quelle que soit la méthode (téléservices, saisie en guichet, capteurs sur le terrain, achat en masse d'informations à un tiers…), d'autant plus s'il est nécessaire de disposer d'informations de qualité (fiable, sans doublon, …), ce qui peut nécessiter vérification, recoupement, retraitement, … ; au-delà du coût pour l'État, ce recueil présente, le cas échéant, **un coût pour celui qui la fournit** (l'usager, l'entreprise). Il vient alors naturellement à l'esprit qu'il est nécessaire d'optimiser ce recueil :

- Eviter de recueillir plusieurs fois les mêmes données. A l'échelle de l'État, cela conduit immédiatement aux initiatives de type « dites le nous une fois », et à chercher à **mettre en commun les données** entre plusieurs administrations, et à choisir des données utiles au plus grand nombre, dans des formats pertinents.
- **Eviter autant que possible les saisies, les ressaisies, les intermédiations**,… et privilégier l'interconnexion avec les systèmes d'information existants **au plus près de la source de référence**. Pour le système d'information de l'État, cela signifie qu'il doit s'ouvrir à ses partenaires au travers notamment d'interfaces « de système à système » et développer une vision du système d'information au-delà de ses frontières de responsabilité.

Pour l'État, la valorisation de cet actif s'entend de nombreuses manières :

- **La donnée doit être disponible et accessible à toutes les administrations** qui en ont opérationnellement besoin, **de façon documentée**, avec des mécanismes de connexion simples à mettre en œuvre (dans un cadre légalement établi lorsqu'il s'agit par exemple de données personnelles),
- La donnée doit être disponible pour être **traitée à des fins décisionnelles**, d'évaluation de politiques publiques.
- Au-delà des administrations et de ses opérateurs, les données qui ne présentent pas de sensibilité particulière quant à leur confidentialité, ont vocation à être **mises à disposition de réutilisateurs tiers (opendata)**.

Bien entendu, la valeur qui peut être créée dépend de nombreux facteurs, notamment de « qualité » : les données sont-elles complètes, exactes, cohérentes,… ? peut-on y accéder facilement, peut-on les mettre à jour facilement ? leur disponibilité, leur intégrité font-elle l'objet d'un engagement ? Cette qualité dépend des **processus de création, de mise à jour, de suppression des données** ; elle

dépend également des **fonctionnalités d'accès mises en œuvre, de leur structure, de leur format** ; elle dépend enfin de la confiance que le destinataire des données peut avoir globalement en y accédant, c'est-à-dire, au-delà des aspects techniques, elle dépend **des engagements de service, de la gouvernance et de l'organisation en place**.

Il est donc essentiel que l'État se dote de règles de gouvernance des données, intégrant des dimensions techniques, fonctionnelles, métiers et organisationnelles, et les applique progressivement dans l'ensemble de son système d'information.

Le présent document, le cadre commun d'architecture des référentiels de données, a pour ambition de fixer un premier corpus de règles, destinées en particulier à la gestion des données de référence et des dispositifs qui les gèrent, appelés « référentiels de données ».

Le chapitre 2 présente les objectifs du cadre, son périmètre d'emploi, son articulation avec le cadre commun d'urbanisation du SI de l'État, son processus interministériel d'élaboration et d'approbation.

Le chapitre 3 explicite un certain nombre de définitions et de notions essentielles, en insistant en particulier :

- Sur la modélisation sémantique, travail conceptuel indispensable à une compréhension partagée entre les acteurs, en s'appuyant autant que possible sur les normalisations existantes et sur des outils de modélisation ;
- Sur les différents critères de qualité de la donnée,
- Sur ce qui est attendu d'un référentiel de données,
- Sur l'articulation avec la démarche d'urbanisation

Le chapitre 4 constitue à proprement parler le corpus de règles, déclinant de façon plus détaillée les principes suivants issus du cadre commun d'urbanisation du SI de l'Etat :

- D1 : les données sont un bien, un actif de l'État, elles doivent être gérées et valorisées en conséquence.
- D2 : les données doivent être standardisées, définies sur la base d'un vocabulaire commun, contextualisées, et combinables les unes aux autres.
- D3 : les données doivent être facilement réutilisables, partageables et accessibles à travers les frontières des administrations.
- D4 : les données publiques doivent être mises à disposition librement et ouvertement sur internet
- D5 : Sécurité et archivage des données

Le chapitre 5 propose quelques architectures logiques types pour la construction de référentiels, à choisir en fonction du contexte. Sans prétendre à l'exhaustivité, ces architectures types ont vocation à couvrir l'essentiel des cas et doivent servir de base aux travaux de conception et d'évolution de tous les référentiels de données de l'État.

Le chapitre 6 donne quelques indications sur la méthode de mise en œuvre progressive de ces travaux dans les différentes administrations.

2. OBJECTIFS DU CADRE COMMUN D'ARCHITECTURE

2.1. Pourquoi un cadre commun d'architecture des référentiels de données ?

2.1.1 Un axe stratégique de transformation du SI de l'État

Le **Cadre Stratégique Commun du SI de l'État**[1] défini les orientations stratégiques de transformation du SI de l'État. Pour organiser et structurer durablement cette transformation, le **Cadre Commun d'Urbanisation du SI de l'État**[2] définit le vocabulaire, les principes applicables, et globalement la démarche d'urbanisation à conduire visant à simplifier, à optimiser, et à rendre durablement plus flexible et agile le système d'information de l'État.

Le présent document n'a pas vocation à reprendre les orientations, objectifs ou principes décrits dans ces deux documents. Il a pour objectif de compléter le cadre commun d'urbanisation du SI de l'État sur le sujet de la gouvernance des données, et en particulier sur les référentiels de données.

La gouvernance des données est l'un des objectifs majeurs de la démarche d'urbanisation ou d'architecture d'entreprise. La mise en place d'une gouvernance consiste en premier lieu à considérer les données manipulées par l'État comme un actif stratégique, et à ce titre assurer leur gestion comme telle : recensement, responsabilité, standardisation, faciliter l'accès, la diffusion, la réutilisation, le partage et l'archivage sécurisé pour en maximiser la valeur. Le terme « données » est à considérer dans un premier temps dans son acceptation la plus large. Il désigne aussi bien des données structurées, semi-structurées, non-structurées, brutes ou agrégées et cela quel que soit la nature, le métier ou le sujet sur lequel porte ces données. Le présent document introduira des distinctions importantes, mais pour comprendre les objectifs visés elles ne sont pas nécessaires.

Le cadre commun d'urbanisation du SI de l'État définit un ensemble de principes. Cinq grands principes sont applicables par tous les acteurs de la transformation du SI de l'État sur la gestion des données.

- **D1** : les données sont un bien, un actif de l'État, elles doivent être gérées et valorisées en conséquence.
- **D2** : les données doivent être standardisées, définies sur la base d'un vocabulaire commun, contextualisées, et combinables les unes aux autres.
- **D3** : les données doivent être facilement réutilisables, partageables et accessibles à travers les frontières des administrations.
- **D4** : les données publiques[3] doivent être mises à disposition librement et ouvertement sur internet.
- **D5** : les données doivent être sécurisées et archivées.

Le lecteur de ce cadre d'architecture est invité à se référer autant que nécessaire au Cadre Commun d'Urbanisation du Système d'Information de l'État. Le choix a été fait dans l'élaboration du présent document d'éviter toute redondance de contenu. Les éléments du cadre stratégique ou du cadre d'urbanisation ne seront donc pas repris dans ce document, mais uniquement cités.

Les principes D1, D2 et D3 sont au cœur de ce cadre d'architecture, ainsi que les principes de conception générale C1 à C6, et les principes de construction des services S1 à S4.

Parmi les données manipulées, échangées, traitées par l'ensemble des acteurs, des activités, des processus, des outils numériques ou non, que composent le système d'information de l'État français, certaines ont des caractéristiques particulières : réutilisation, duplication, transversalité, valeur (notamment par rapport aux processus métiers), impact. Le concept de « données de référence », ou « *master data* » est utilisée pour désigner ce type de données. Le terme « référentiels de données » désigne les outils informatiques nécessaires à la gestion de ces données dans le temps et leurs mises à disposition des autres applications, systèmes d'information ou utilisateurs. Les référentiels de données sont des applications clés pour l'ensemble du SI de l'État, et souvent même pour des acteurs externes à l'État. **Ils sont la pierre angulaire de toute la démarche d'urbanisation du SI de l'État**. L'efficacité, la qualité, la pérennité et l'agilité de telles

[1] Le document est disponible en ligne dans sa version 1.0 de février 2013, avec la circulaire du Premier Ministre du 7 mars 2013 : https://references.modernisation.gouv.fr/strategie-du-si-de-letat

[2] Le document est disponible en ligne dans sa version 1.0 de novembre 2012 : https://references.modernisation.gouv.fr/urbanisation-du-systeme-dinformation-de-letat

[3] Le terme « données publiques » fait référence à la circulaire du 26 mai 2011 du Premier Ministre : http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000024072788

applications est un objectif majeur des travaux d'urbanisation du SI de l'État, raison pour laquelle un cadre spécifique leur est consacrées.

Le présent cadre porte sur la construction, l'entretien et l'évolution de dispositifs appelés « référentiels de données » assurant la collecte, la gestion, l'archivage et la mise à disposition de « données de référence » à l'ensemble du système d'information de l'État.

Le présent document est avant tout la consolidation de bonnes pratiques actuelles. Il n'est pas porteur de rupture en soit, mais rappelle en quoi les référentiels de données sont à la fois la pierre angulaire du SI de l'État et de son efficacité et son efficience, et un élément de simplification de la complexité intrinsèque au SI de l'État.

2.1.2 Constats partagés

Les Systèmes d'information des différents ministères, administrations ou établissements publics ont été construits le plus souvent de manière indépendante les uns des autres, et souvent sans concertation, alors que de nombreuses informations sont échangées et manipulées par et entre ces différents acteurs. Il est ainsi fréquent de rencontrer la même information (ex. Usager, agent, structure, entreprise, adresse, etc.) plusieurs fois dans le SI d'un des ministères ou d'administrations et ou encore quand on analyse l'ensemble du SI de l'État. Ces données ont des structures techniques souvent différentes, alors que leur sémantique est identique ou proche, et que bien souvent l'origine de la donnée est la même (ex. l'usager). Les informations sont saisies plusieurs fois dans diverses applications ; ce qui génère doublons, incohérences, donc surcoûts et globalement une inefficacité qu'il est de plus difficile à réduire.

D'autres constats alimentent la nécessité d'encadrer le sujet des référentiels de données :

- Les fonctionnalités et les applications de gestion de données de référence sont développées plusieurs fois, et bien souvent sont incomplètes ou obsolètes.
- La fraîcheur et la qualité de la donnée ne sont pas égales partout dans le SI. La fréquence de rafraîchissement et la qualité d'acquisition des données dépendent bien souvent d'un processus d'acquisition locale et du niveau d'exigence du métier commanditaire, alors que, comme nous le verrons plus loin, une donnée de référence a la caractéristique d'avoir une durée de vie qui va bien au-delà de processus métiers locaux.
- La mise en œuvre d'un nouveau projet est souvent l'occasion de questionnements pour identifier la source des données de référence. L'analyse n'est souvent pas encadrée globalement, et la solution trop souvent locale compte tenu de contraintes légitimes de projets.
- Dans le cas de processus transverses à plusieurs SI, les applications sont obligées de procéder à des trans-codifications entre leur données de référence et les données d'un autre SI Métier, avec tous les risques d'erreurs quand le cycle de vie de ces données n'est pas parfaitement encadré (acquisition, validation mise à jour, conservation, publication, suppression logique, archivage, etc.).
- La mise en place d'un SI Décisionnel est rendue difficile, puisqu'il ne dispose pas, ou peu de données transverses de bonne qualité pour constituer ses axes d'analyse et de segmentation pertinent pour le pilotage de l'action publique. Ces outils de pilotage mettent en évidence, en bout de chaîne, le niveau de qualité des données.
- La mise en place de processus transverses est difficile puisque d'un SI à un autre les informations peuvent être structurées et gérées de manière différentes alors que leur sémantique est bien souvent identique.

Ces constats font apparaître le besoin d'un cadre, d'une approche voire de solution pour disposer :

- d'une vision unique de la donnée de référence partagée (Sémantique, structure, méta donnée..), sous la responsabilité d'un acteur métier reconnu.
- du juste niveau de qualité des données de référence pour l'ensemble des utilisations dans le SI de l'État.
- du catalogue et de la roadmap des référentiels de données et leurs services d'accès aux données.

D'où le recours à la mise en place de dispositif spécifique, les référentiels de données, qui en résumé, doivent permettre :

- d'améliorer la productivité de l'organisation : En utilisant un vocabulaire commun pour les processus de bout en bout, un référentiel facilite la communication entre les différents métiers.
- de permettre des économies d'échelle : Les économies sont réalisées en réduisant les coûts supplémentaires et inutiles de gestion de référentiels redondants (principe de mutualisation). Il évite également la gestion des erreurs coûteuses en raison des différences locales et des incohérences dans les données de référence.
- d'augmenter la qualité de l'information : La mise en place d'un référentiel unique, permet le maintien d'un bon niveau de qualité (fraîcheur et cohérence) dans les données de référence. C'est un facteur clé de succès pour améliorer la qualité globale de l'information.
- de simplifier l'audit et la traçabilité : La centralisation des données communes dans les référentiels rend la comparaison plus facile et permet des fonctions transverses telles que le décisionnel de trouver des vues agrégées.

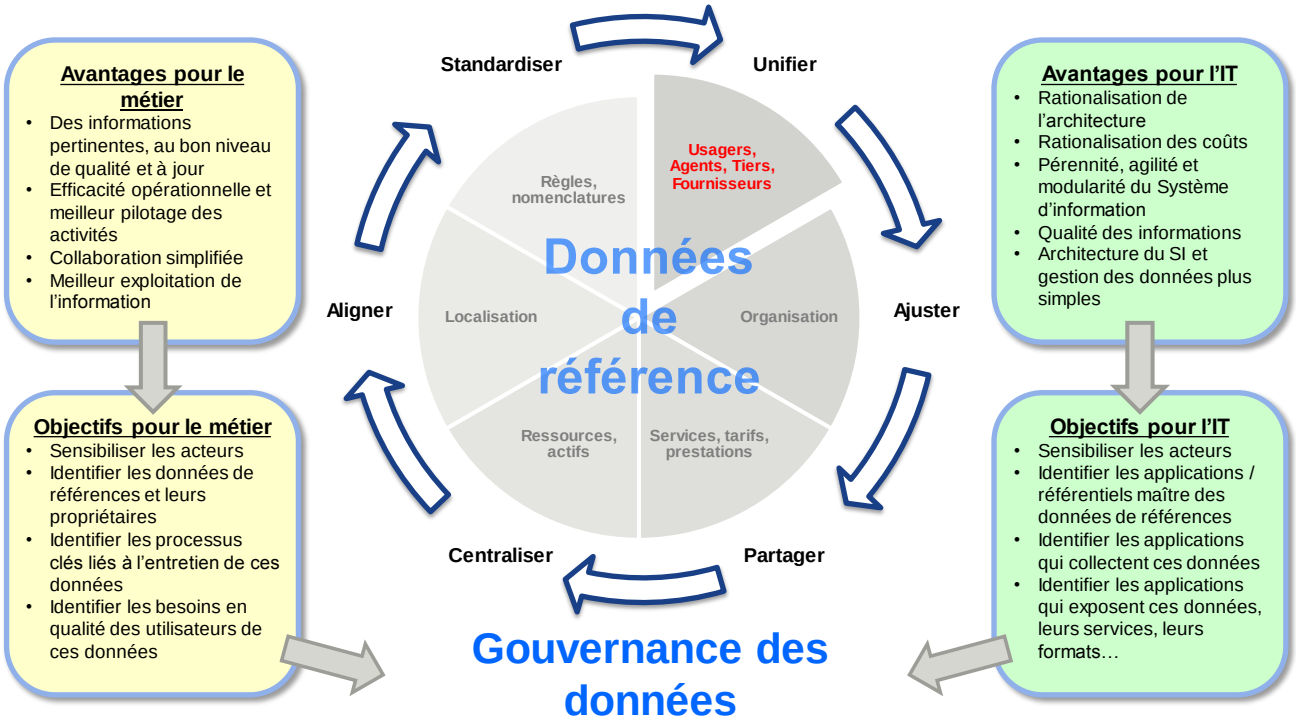

Figure 1 - Avantages et objectifs d'une gouvernance des données

Le présent document donne des règles de constructions et d'entretiens de tels dispositifs. Il est toutefois très important de préciser un point : l'architecture globale de gestion et de distribution de données au sein d'un système d'information comme celui de l'État est une tâche particulièrement complexe qu'il convient de traiter avec l'attention et la coopération de tous les acteurs.

2.1.3 Objectifs

Le présent document a pour objectif majeur d'initier la mise en place au niveau de l'État, d'une gouvernance de données. Il a été décidé d'amorcer cette mise en place par un premier travail sur les données de référence et les référentiels de données. Ce présent Cadre commun d'architecture des référentiels de données a donc pour objectifs :

Objectif 1 : Fournir un langage commun, un cadre pour tous les acteurs engagés dans l'entretien, la transformation et l'évolution des référentiels de données de l'État, et **améliorer la coopération et l'efficacité de ces acteurs.**

Objectif 2 : Rappeler pourquoi la mise en place de référentiel de données est un élément clé d'une architecture de système d'information.

Objectif 3 : Positionner l'analyse sémantique comme une démarche clé dans l'identification, la compréhension partagée et la structuration des référentiels de données.

Objectif 4 : Augmenter la qualité et l'auditabilité des données de référence globalement dans le SI de l'État. Même si c'est objectif vise le plus long terme, c'est très clairement l'objectif à atteindre.

Objectif 5 : Définir les règles d'architecture, c'est-à-dire de construction et d'entretien des référentiels de données, et les critères de succès de mise en place de ces référentiels : des règles communes à tous alignées sur les meilleures pratiques du moment.

Objectif 6 : Aider à l'évaluation des référentiels existants par rapport à ces règles de construction et d'entretien : le présent document en effet propose un cadre commun pour le diagnostic et l'analyse des référentiels existants au sein du SI de l'État.

Objectif 7 : Faciliter la transformation de ces référentiels dans le temps : mutualisations, simplifications durable, rationalisations, évolutions pour plus d'efficacité, de qualité, d'agilité.

Objectif 8 : Identifier formellement et labelliser les référentiels de données du SI de l'État : l'identification, ou la mise en place le cas échéant, des référentiels de données pour l'ensemble des données transverses, est un facteur de succès de ce travail.

2.2. Portée du cadre d'architecture

Ce cadre s'adresse à tous les acteurs de la transformation du SI de l'État, et en particulier sur le périmètre de l'Administration de l'État. Il vise en particulier les maîtrises d'ouvrage (MOA) stratégiques des ministères, les directeurs et chefs de projets MOA, les responsables des services des archives, les DSI et leurs comités de pilotage, l'ensemble des urbanistes et architectes SI, des architectes techniques, ainsi que les directeurs et chefs de projet MOE.

La gouvernance des données n'est absolument pas qu'un problème technique et d'informatique, elle n'est pas le fait de quelques acteurs ou experts, même si elle peut être impulsée ou coordonnée par quelques-uns, mais bien l'affaire de tous, en particulier les directions métiers qui doivent comprendre et porter l'intérêt d'une gouvernance sur l'élément qui est au cœur du Système d'Information de l'État : les données.

L'urbaniste SI a avant tout un rôle de catalyseur, de facilitateur et de coordinateur dans la démarche.

Ce document ne se veut pas un support de formation à la mise en place et la gestion de référentiels de données (qui est le dispositif technique), ou la gestion de données de référence (qui est le dispositif métier plus global). Le chapitre sur les règles, notamment, ne présente volontairement pas tous les éléments justificatifs. Il s'agit avant tout de consolider, structurer, parfois recadrer les pratiques existantes, depuis des années pour certaines. Les lecteurs qui souhaiteraient approfondir telles ou telles parties du présent cadre trouveront des réponses dans les documents cités en références et se rapprocheront des Urbanistes SI des ministères.

2.3. Articulation avec le corpus réglementaire

Le présent cadre d'architecture constitue un des éléments du corpus réglementaire applicable pour la construction, la gestion, l'exploitation et la transformation du SI de l'État, à tous les niveaux. La figure ci-après illustre les différents documents applicables et leurs articulations (répartis selon les vues du SI).

Ce corpus se compose de documents de politique globale applicables à l'ensemble du SI de l'État :

- le Cadre stratégique, qui définit la stratégie de l'État en matière de SI[4],
- la Politique de Sécurité, définissant les règles générales de sécurité[5],
- le présent Cadre commun d'urbanisation définissant la démarche d'urbanisation,

Ce corpus comprend également des documents réglementaires techniques à portée plus large (administrations de l'État, collectivités territoriales, organismes de la sphère sécurité et protection sociale), définis par l'ordonnance n°2005-1516 du 8 décembre 2005 :

- le Référentiel Général d'Interopérabilité[6],
- le Référentiel Général de Sécurité[7],
- le Référentiel Général d'Accessibilité pour l'Administration[8].

Il comprend également des documents à portée ministérielle :

Un cadre stratégique ministériel (ou schéma directeur), déclinant le cadre stratégique du SI de l'État, dans le contexte métier d'un ministère.

- un cadre de cohérence technique : normes, standards et règles d'architecture applicables localement ;
- une méthode de conduite de projet qui définit et structure les relations MOA/MOE et le pilotage des projets de transformation du SI pour un ministère (ou une administration) ;
- éventuellement une charte d'urbanisation qui décline localement le cadre commun d'urbanisation, sans pour autant modifier les principes, ou le cadre d'activité, mais uniquement en précisant l'organisation, les méthodes de travail, et le fonctionnement spécifique à chaque ministère.

[4] Se référer à l'article 4 du décret n°2011-193 du 21 février 2011 concernant la création de la DISIC
[5] Se référer aux Politiques de Sécurité ministérielles et à la Politique de Sécurité du SI de l'État.
[6] http://references.modernisation.gouv.fr/rgi-interoperabilite
[7] http://references.modernisation.gouv.fr/rgs-securite
[8] http://references.modernisation.gouv.fr/rgaa-accessibilite

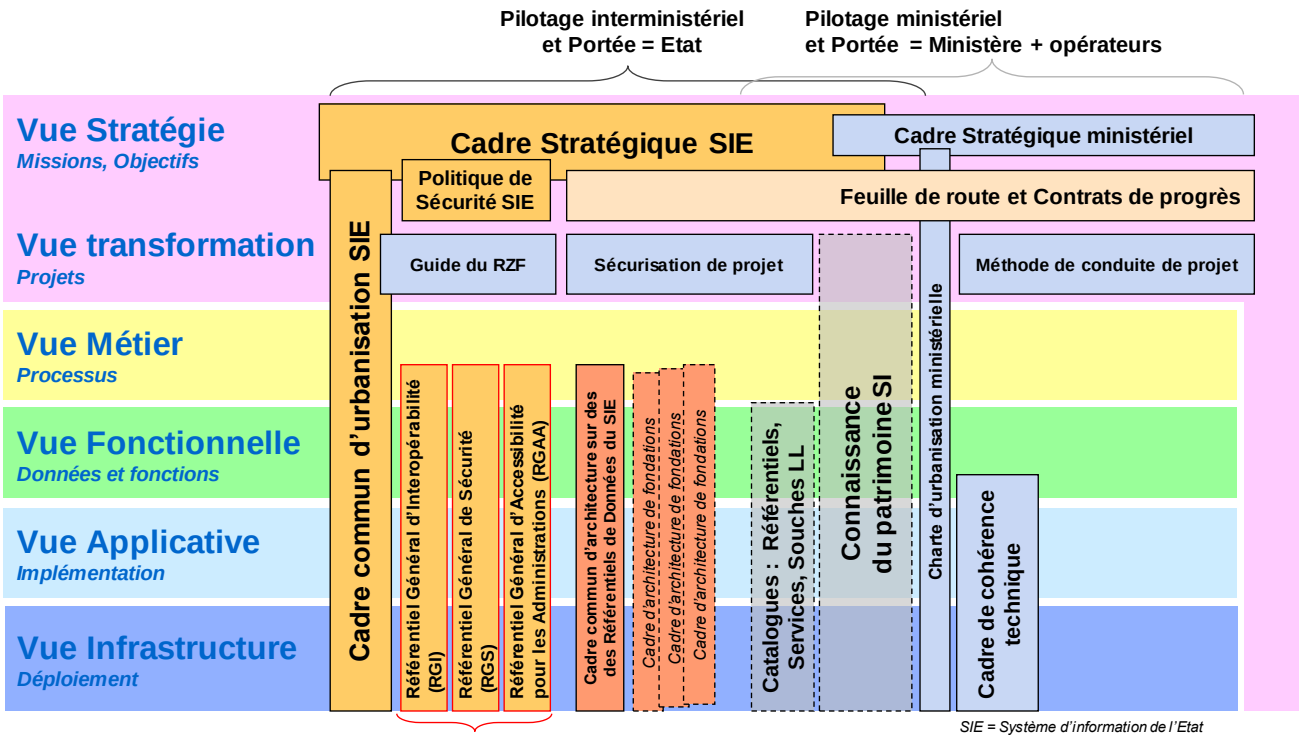

Figure 2 - Articulation du Cadre Commun d'Architecture des Référentiels de données avec l'ensemble du corpus réglementaire

Ce corpus pourra également comprendre des cadres d'architecture sur les fondations du SI de l'État à portée globale. Les sujets de la gestion des identités, des accréditations et des accès au SI de l'État, mais aussi de gestion des référentiels de données, du décisionnel, etc. sont tous des sujets complexes et transverses qui nécessitent d'être cadrés et structurés par un cadre d'architecture spécifique et à vocation interministérielle.

2.4. Entretien et mise à jour du document

Toutes les demandes d'évolution devront être adressées à la DISIC directement, ou via un des urbanistes membres de la communauté interministérielle qui proposera une adaptation du cadre, ou bien sûr via un responsable de secteur fonctionnel du domaine « données transverses ». Ces adaptations seront analysées lors des ateliers de travail interministériels avec les métiers en charge des référentiels, puis soumis à validation au comité technique des SIC (CTSIC).

2.5. Version du document et références documentaires

Ce document est le résultat d'un travail collectif de la communauté interministérielle des urbanistes, représentant l'ensemble des ministères et animé par la DISIC. Il est le produit de l'analyse des besoins et des objectifs des différents ministères, de l'expérience de chacun des membres de cette communauté, ainsi que des meilleures pratiques du moment.

Il complète le cadre d'urbanisation :

- ***Cadre Commun d'Urbanisation du Système d'Information de l'État***[9] version 1.0 du du 26/10/2012.

Le présent cadre s'inspire et reprend également des éléments d'ouvrages, de cadres ou de documents de natures équivalentes, de ministères, d'opérateurs, d'autres pays, ou d'associations internationales :

- ***Directive sur l'Administration des données*** n°14/DEF/DGSIC du 19/07/2010 du Ministère de la Défense ;
- ***Urbanisation – Gestion de données de référence***, version 1.1 du 10/01/2013, DSIC, Ministère de l'intérieur ;
- ***Les référentiels dans un système d'information***, du 16/11/2010, CNAV ;
- ***Urbanisation, SOA et BPM***, Yves Caseau, DUNOD, 4ième édition 2011 ;
- ***Australian Government Architecture Reference Models***[10] version 3.0 d'août 2011 (Australie) ;
- ***Queensland Governement Enteprise Architecture Framework***, version 2 de 2009, État du Queensland (Australie), et en particulier les documents relatifs à la gouvernance de l'information ;

[9] http://references.modernisation.gouv.fr/urbanisation-du-systeme-dinformation-de-letat
[10] http://www.finance.gov.au/e-government/strategy-and-governance/australian-government-architecture.html

- ***Les référentiels du Système d'information***, Joël Bizingre, Joseph Paumier, et Pascal Rivière, DUNOD, 2013 ; **Cet ouvrage est l'un des plus complets et plus riches écrit à ce jour sur ce sujet**. Il a très significativement inspiré le présent cadre d'architecture.
- ***MDM***, Franc Régnier-Pécastaing, Michel Gabassi et Jacques Finet, DUNOD, 2008 ;
- ***Enterprise Data Governance***, Pierre Bonnet, ISTE - Wiley, 2010 ;
- ***Guide de l'aspect sémantique de la Méthode PRAXEME***, version 2, 2013, Praxeme institute ;
- ***Architecture des SI, Livre Blanc***, OCTO Technologie, 2002.

Le tableau ci-après résume l'évolution du cadre avec les principales versions :

Version	Date	Motifs
0.1	12/03/2012	Première version de travail du document pour initier l'élaboration du Cadre
0.6	01/05/2013	Version intermédiaire de travail
0.8	17/06/2013	Version intermédiaire de travail
0.9	24/07/2013	Premier complément et version complète du document pour diffusion large

Tableau - Suivi des versions du Cadre

Le tableau suivant cite le nom des personnes de la communauté des urbanistes SI qui ont directement contribué à l'élaboration et la rédaction de cette première version du Cadre Commun d'Urbanisation du SI de l'État :

Nom, Prénom	Ministère
BANAT-BERGER Françoise	Archives de France (SIAF)
BARBAY Alain	Intérieur
BOURDIN Didier	Économie, Finances, Fonction Publique / Opérateur National de Paye.
CAFFIERY Geoffroi (LCL)	Défense
HEIJLIGERS Éric	Justice
JIMENEZ Francisco	Affaires Étrangères
LE BRAS Maryse	Éducation Nationale, Enseignement Supérieur et Recherche
LASFER Leila	Agriculture
MAATOUG Ridha	Culture et Communication
MILLEFAUX Laure	Écologie, Développement Durable, Énergie
MUSET Catherine	Éducation Nationale, Enseignement Supérieur et Recherche
HERICHER Benoît	Intérieur
PETITIMBERT Stéphane	Économie, Finances, Commerce extérieur, Redressement productif, Réforme de l'État, Décentralisation et Fonction Publique.
PIERRE-DIT-MERY Luc	Premier ministre / DISIC (**rédacteur**)
REGNIER Jérôme	Affaires Sociales et Santé
RIVIERE Pascal	Affaires Sociales et Santé (CNAV)
SOUSSAN Claude	Écologie, Développement Durable, Énergie
VERCAUTEREN Pierre	Travail, Emploi, Formation Professionnelle et Dialogue Social

Figure 3 - Tableau des contributeurs à la rédaction du Cadre Commun d'Urbanisation du SI de l'État

2.6. Diffusion

Ce cadre d'architecture est destiné à une diffusion publique et ouverte de manière à atteindre l'ensemble des communautés concernées par l'entretien et les évolutions du Système d'Information de l'État, et en particulier des référentiels de données qui le composent. Il est également destiné aux prestataires et partenaires de l'État. Il sera accessible sur le site web de la DISIC, et sur l'outil collaboratif utilisé par les communautés interministérielles (notamment « Urbanisation » et « Interopérabilité »).

Les Ministères auront la charge de le décliner opérationnellement, avec l'appui et l'animation de la DISIC dans le cadre stratégique commun du SI de l'État.

2.7. Approbation

Le présent document fera l'objet d'une validation d'une part, par la communauté des Urbanistes SI de l'État, et d'autre part, par le dispositif de gouvernance mis en place par la DISIC à savoir le Comité Technique des SIC (cf. décret n° 2011-193 de création de la DISIC du 21 février 2011).

3. DÉFINITIONS ET PRINCIPES

La mise en place partielle ou en totalité d'une gouvernance de données passe en premier lieu par une clarification et un partage des concepts manipulés. Dans tous les travaux relatifs aux données, la question de la sémantique est au cœur des réflexions. Ce présent cadre n'échappe pas à cette règle. Il est donc utile et nécessaire d'expliquer et de définir le sujet traiter, en particulier :

- qu'est-ce qu'une donnée ?
- qu'est-ce qu'une donnée de référence ou transverse ?
- qu'est-ce qu'un référentiel de données ?

Le présent chapitre définit les concepts clés liés à la mise en place, l'utilisation ou la transformation de référentiel de données. Tous les concepts sous-jacents du présent cadre sont définis en annexe.

3.1. Une donnée

Une donnée est une description élémentaire de nature numérique, représentée sous forme codée, d'une réalité (chose, évènement, mesure, transaction, etc…) **en vue d'être** :

- **collecté**, enregistrée,
- **traitée**, manipulée, transformée
- **conservée**, archivée
- **échangée**, diffusée, communiquée.

Il peut être question de **donnée structurée**, **semi-structurée** ou **non-structurée**.

Une donnée structurée est une donnée dont on a établi fonctionnellement le sens de manière détaillée, son cycle de vie et donc ses règles de création, les valeurs possibles dans le cas de listes, ainsi que le moyen technique de représentation. Les données structurées sont stockées ou échangées dans un format organisé, une syntaxe, et dont la sémantique a préalablement été définie (par exemple une table dans une base de données, un document XML…).

Les données non-structurées sont également une description d'une réalité, mais dont la codification, dont le sens, n'est pas exploitable directement par la machine : un fichier audio, vidéo, un texte contenu dans un document ou un mail par exemple. Des données semi-structurées combinent les deux types : une partie structurée, et une partie non structurée.

Une **information** est un ensemble de données agrégées en vue d'une utilisation par l'homme.

3.2. Sémantique et modèle sémantique

3.2.1 Définir pour comprendre et se comprendre

Appliquée au domaine qui nous intéresse, **la sémantique est l'étude et la formalisation du sens et de la signification des données**. Ce n'est absolument pas une problématique technique mais une discipline réellement métier. Il s'agit de formaliser, de structurer la compréhension d'une réalité, qui se traduit en données.

Quels sont les notions, concepts et objets du domaine étudié ? Quel est le cycle de vie de ces objets ? Quels sont les comportements de ces objets, de ces réalités ? Quelles relations ont-ils entre eux ? Quelles sont les règles qui les contraignent ?

Les technologies de l'information et de communication ont un apport indéniable dans le fonctionnement et l'efficacité de l'État. L'objet du propos n'est pas ici de l'expliquer. Ces technologies permettent notamment d'échanger des données. C'est précisément parce qu'il y a eu une entente entre les acteurs sur la sémantique de ces données, sur leur sens, que ces échanges sont possibles, et permettent à ces acteurs d'opérer de concert, avec une efficacité très souvent liée à la richesse sémantique et la qualité des données échangées. Ce travail est bien souvent implicite car il est évident pour tous.

Pour une plus grande efficacité sur le long terme, il est nécessaire de créer et d'entretenir un dictionnaire sémantique des données les plus échangées. C'est-à-dire un document, ou un ensemble de document, décrivant la sémantique de ces données.

Prenons un exemple : le cas d'un échange de données concernant l'adresse d'établissements (localisation d'une structure territoriale d'une administration par exemple). Au premier abord, la compréhension est assez évidente sur ce qu'est une adresse : dans notre cas, c'est un ensemble de données permettant de localiser une organisation. Il faut toutefois noter que

dans un certain nombre de pays, l'adresse n'est pas aussi formalisée et structurée qu'en France. Et dans notre exemple, si l'échange doit s'opérer entre deux administrations, il est probable que la réflexion sur la sémantique soit rapide et que les deux parties s'entendent sur la sémantique de ces données, pour passer rapidement à la réalisation des outils informatiques nécessaires à l'échange.

Dans le cas d'une réflexion plus globale, au niveau de l'État Français, sur des données de référence, la question sur la sémantique prend une tout autre dimension.

Reprenons notre exemple sur l'adresse d'établissements : s'agit-il de disposer d'une adresse postale pour des envois de courrier ? D'une adresse d'accès physique précise jusqu'au bâtiment voire au bureau pour identifier par exemple les établissements qui sont ouvert au public (avec des informations sur les conditions d'accès) ? D'une adresse de livraison pour acheminer des ressources particulières, ou encore d'une adresse d'accès pour les services d'urgence ? Il est vrai que dans de nombreux cas, ces différentes adresses se confondent, mais ce n'est pas une généralité (cas des grands sites accueillant plusieurs structures de l'État, et disposant de plusieurs accès physique : par exemple une préfecture). A l'échelle de l'État, le nombre de cas de ce type ne peut pas être simplement ignoré. De plus, toujours dans notre exemple, Le groupe La poste a mis en place des dispositifs pour ces besoins propre de distribution du courrier (par exemple les CEDEX) qui rajoutent une forme de complexité. Par ailleurs que se passe-t-il dans notre échange, quand une rue est renommée, ou renumérotée ? Que se passe-t-il si l'adresse est erronée ? Si une zone est en travaux (rénovation d'un quartier) ? Ne serait-il pas plus simple de gérer des coordonnées géographique de type GPS ? Encore une fois à l'échelle de l'ensemble du Pays, ce ne sont pas des cas isolés.

De plus, concernant les données sur les établissements que l'on souhaite échanger : de quoi s'agit-il finalement ? Comment ces établissements sont-ils identifiés ? Par leur SIRET ? Sont-ils liés entre eux et comment ? Que se passe-t-il dans les cas de réorganisations d'une administration ou d'un réseau territorial ?

Un autre élément fondateur dans ce travail sur la sémantique est la nécessité d'élargir le point de vue. Formaliser le sens ou la signification de données ne peut pas être fait uniquement par rapport à l'usage de ces données, mais bien par rapport à la réalité que ces données décrives. Il est donc primordial de se poser des questions sur l'origine de ces données et sur les processus qui les ont créées. Prenons un exemple avec les données sur les entreprises utilisées par la DGFiP. Ces données sont souvent qualifiées de « données fiscales » sur les entreprises. C'est un abus de langage, car ce terme qualifie l'usage qu'il est fait de ces données, mais pas leur sémantique, leur sens intrinsèque. Il est nécessaire de comprendre l'origine de ces données au sein de l'entreprise, et les réalités qu'elles décrivent. Dans notre exemple, l'origine de ces données est la plus part du temps est la comptabilité de l'entreprise, mais pas uniquement, une autre partie de ces données proviennent également des ressources humaines (la paye notamment).

3.2.2 Monter en abstraction pour simplifier, rendre plus robuste et flexible

Là encore on comprend aisément l'importance de ces questions dont la réponse ne peut être apportée que par les acteurs métiers. Elles prennent une dimension particulière dans le cadre de l'automatisation de leur gestion par des outils informatiques, mais cela ne change en rien le fait qu'elles adressent très clairement un savoir-faire métier, et parlent donc de fait, aux utilisateurs, et leurs responsables, aux juristes et donc leurs maîtrises d'ouvrage.

L'étude de la sémantique des données, c'est-à-dire l'étude de la sémantique des objets et des concepts qui sont au cœur de l'activité, doit se débarrasser le plus possible des contingences organisationnelles et techniques. Il s'agit de décrire le plus fidèlement possible la réalité des objets manipulés. La valeur ajoutée de ce travail réside dans la capacité à monter en abstraction pour finalement en extraire ce qui est stable et simple. C'est très précisément l'abstraction sur ce travail de sémantique qui permet de disposer d'une description simple, robuste, générique et donc plus agile, modulaire, efficace et donc plus efficiente sur le long terme. Cette étude de la sémantique s'appuie également sur les préceptes de l'approche orientée objet, qui pousse à la généricité, et permet d'établir un lien plus précis avec la conception, et le développement des outils logiciels.

Un **objet métier** est un concept ou une abstraction ayant un sens pour des parties prenantes interne ou externe d'une organisation. L'objet métier permet de décrire les données manipulées, dans le cadre d'exécution de processus, de projet ou de tâche ad hoc, par des organisations, des personnes, des applications, des systèmes… L'objet métier peut être matériel (exemple, tout produit reçu ou expédié, une table, un wagon, etc.), immatériel (exemple, un service, un compte bancaire, etc.), ou encore virtuel (exemple, une réunion, un service d'organisation, etc.). Un objet métier se caractérise par un **état**, il est représenté par des **données d'identité**, des données descriptives ou **des attributs**, **des comportements**, et **des relations** (ou encore **des associations**, le terme est également utilisée) qu'il entretient avec les autres objets.

Un objet métier est donc une abstraction d'un ensemble cohérent de données qui a un sens pour un ou plusieurs acteurs : par exemple, une personne, un contrat, une entreprise. Ces concepts, ces objets métiers donc, ont bien un sens pour de nombreux acteurs au sein de l'administration, et ils constituent chacun un ensemble de données cohérent.

Attention, car c'est une erreur fréquente et aux conséquences financières parfois considérables : ce n'est pas en simplifiant ou en éludant certaines questions et réponses sur l'étude de la sémantique des objets métiers qu'il est possible d'obtenir plus rapidement un modèle sémantique plus abstrait et robuste. C'est tout le contraire. L'abstraction sur des questions de sémantique consiste à identifier les définitions, les caractéristiques, les comportements communs entre des

objets et des concepts qui ne le sont pas a priori. C'est ce travail qui permet d'obtenir une représentation plus simple et robuste. Cela n'en est pas moins concret et précis, même si l'abstraction, présente un inconvénient qu'il convient de savoir évaluer : la lisibilité par l'ensemble des acteurs concernés indépendamment de leur niveau de maturité sur ce sujet.

Le moindre défaut sur cette analyse sémantique aura donc un impact fort sur le long terme. D'où un autre facteur de succès important : la standardisation.

3.2.3 Standardiser pour l'interopérabilité

La recherche de la standardisation sur ces questions sémantiques est fondamentale. Il existe de nombreux standards en la matière. Les principes D2 et C2 du cadre commun d'urbanisation du SI de l'État précisent pourquoi cette recherche de la standardisation est critique. C'est une garantie de pérennité et d'acceptation par le plus grand nombre. L'UN/CEFACT[11] publie une liste de description sémantique d'un certain nombre d'objet courant et transverse, *Core Concept Library (CCL)* : une personne, un contrat, une adresse, une organisation, une identité, etc. Il en existe de nombreuses autres. L'objectif du Référentiel Général d'Interopérabilité est de référencer les descriptions sémantiques retenus et valides pour l'ensemble de l'État (collectivités territoriales comprises). Le Modèle de Données Communes (MDC) annexé au RGI v1.0 a été construit très précisément à partir de l'UN/CEFACT. C'est très clairement un axe d'effort pour la prochaine version du RGI.

3.2.4 Formaliser et modéliser

Cette sémantique des données doit être formalisée pour être aisément communiquée, partagée, pérennisée et réutilisée (notamment par les équipes en charge de la conception et du développement des applications : il est question de modélisation, mais pas uniquement de modélisation de données, sous la forme de diagrammes de classe, mais également de modélisation du cycle de vie des données, sous la forme de diagrammes d'état.

La description de la structure des concepts manipulés est tout aussi importante que la description de la dynamique de ces concepts (états, évènements…), et la description des règles qui régissent ces objets

Autant la réflexion de fond sur la sémantique est du ressort d'acteurs métiers (MOA, AMOA), autant la représentation sous forme de modèle respectant une notation normalisée nécessitent une compétence très spécifique et donc bien souvent l'intervention d'acteurs de DSI (urbaniste, architecte, concepteur / modélisateur). Ces travaux sont donc nécessairement le fruit de la coopération de ces 2 familles d'acteurs.

L'utilisation de modèle, et l'apport de la modélisation n'est plus à démontrer aujourd'hui, même si les compétences en la matière au sein de l'État sont plutôt rares. Il n'y a que par cette modélisation, cette représentation graphique synthétique et organisée, que l'on réduit les risques d'erreur d'interprétation de la sémantique. Un modèle présente également deux autres avantages par rapport à une description textuelle :

- il est plus compact, c'est-à-dire qu'il véhicule beaucoup plus d'information que du texte sur une même surface,
- et il est régit par des règles de construction et de lecture, qui imposent une rigueur dans sa construction.

Par construction, un modèle est un outil de conception, mais aussi un instrument de communication entre les acteurs concernés. D'où l'utilisation d'une représentation formelle et standardisée. **La norme UML12 (ISO/IEC 19505-1 et 19505-2), est désormais un incontournable. Tous les travaux de modélisation de sémantique devront s'appuyer sur cette notation**. Trois diagrammes en particulier seront utilisés : les diagrammes de classes pour la représentation statique, ils pourront être accompagnés d'exemple sous la forme de diagramme d'objets ; les diagrammes de machine à état (ou diagramme d'état) pour la représentation dynamique du comportement des objets métiers ; les diagrammes de package pour la représentation des ensembles cohérents d'objets métiers.

Il existe par ailleurs, des représentations simplifiées, en cours de construction. Des travaux internationaux sont en cours rapprochant les différentes approches issues des mondes « objet », « relationnel » et « web ».

Globalement, l'ingénierie dirigée par les modèles, et plus particulièrement l'architecture dirigée par les modèles (MDA[13]) pousse la logique de modélisation en proposant des outils pour la déclinaison de modèles métiers indépendant de leur informatisation, en modèles informatiques indépendant de la plate-forme technologique retenue, puis de les décliner à leur tour en modèle dépendant de chaque plate-forme / technologie utilisée.

[11] UN/CEFACT : *United Nations Centre for Trade Facilitation and Electronic Business*, propose des recommandations, des standards et des spécification techniques pour faciliter les échanges électroniques pour le commerce international.
http://www.unece.org/cefact.html

[12] UML : *Unified Modeling Language*, langage de modélisation graphique proposée et soutenue par l'Object Management Group (OMG), et normalisé au niveau ISO

[13] MDA : *Model Driven Architecture*, démarche d'architecture proposée et soutenue par *l'Object Management Group* (OMG)

3.3. Les données de référence

Parmi les données collectées, traitées, manipulées, ou échangées au sein du SI de l'État, certaines ont des caractéristiques particulières, au nombre de cinq : il est question alors de **données de référence**. Les cinq caractéristiques principales des données de référence sont les suivantes :

1. **elles sont utilisées fréquemment par un grand nombre d'acteurs internes ou externes** (organisations, métiers, processus, applications…). Elles peuvent être utilisées par des métiers fondamentalement différents. Par exemple certaines données sur les entreprises sont utilisées aussi bien par les sphères fiscale, sociale, travail, emploi, développement durable, santé, agriculture.
2. **leur qualité est critique pour un grand nombre de processus**. Elle conditionne directement l'efficacité et l'efficience de ces processus, et donc plus globalement impacte le pilotage de l'action publique.
3. **leur sémantique est partagée et relativement stable dans le temps**. L'unicité et la richesse sémantique de ces données est recherchée pour simplifier les processus, optimiser leurs exécutions, et apporter plus de valeur aux clients de ces processus. La portée de ces données, c'est-à-dire la couverture d'usage de ces données, est également un critère clé dans leurs utilisations, et des incompréhensions sur cette portée peuvent impacter également l'efficacité des processus.
4. **Elles ont une durée de vie qui va au-delà des processus opérationnels qui l'utilisent.** De fait, les données de contextualisation qui leurs sont associées, c'est-à-dire leurs métadonnées, sont critiques.
5. **La facilité d'accès à ces données est critique et conditionne l'efficacité et l'efficience global** des solutions mise en place pour utiliser / exploiter ces données : depuis n'importe où, tout le temps, et quel que soit le dispositif technique qui en a besoin. L'identification des données de référence est un sujet particulièrement sensible et conditionne l'efficacité des échanges et de l'exploitation de ces données (identifiant unique et partagé). L'interopérabilité des dispositifs d'accès à ces données est une condition de succès.

Note : Les données de référence sont de faite des données structurées, ou semi-structurées.

En résumé, les caractéristiques principales d'une donnée de référence sont donc le sens (la sémantique), la qualité, le partage et la réutilisation (en consultation principalement, et donc sans modification) par plusieurs acteurs et applications du système d'information.

Il est fréquent de distinguer trois grands types de données de référence :

- les données « maître », qui sont en général les objets métiers principaux d'un domaine fonctionnel (Entreprise, Personne, Structure, Agent), et qui donc correspondent aux principales zones du domaine « Données transverses » du POS du SI de l'Etat ;
- les données « constitutives », qui caractérisent en général les données maître ou les complètes, mais aussi d'autres objets métier (par exemple : les moyens de contacts et de paiement d'une personne ou d'une entreprise ; les données comptables d'une entreprise, les données de revenus d'une personne) peuvent caractériser des données maître mais également d'autres objets métiers liés ;
- les données « paramètres » ou tables de valeurs, ou encore nomenclatures (par exemple codes postaux, code banque, codes devises, grades, taux de taxes, …), ce sont les données les plus partagées au sein du SI. Elles servent à indexer, classifier, organiser, structurer, hiérarchiser l'information. Ces nomenclatures peuvent prendre différentes formes ; il est question de simple plan de classement, ou encore de thesaurus, d'ontologie, de folksonomie.

Il convient toutefois de souligner que la nuance entre donnée « maître », « constitutive », « paramètre » dépend du périmètre d'analyse, et de l'acteur qui procède à l'analyse. Une donnée constitutive pourra être considérée comme maîtresse pour un métier, car elle se trouve au cœur de ses processus (exemple les données concernant les professionnels de santé, peuvent être considérées comme « maîtres » pour plusieurs domaines fonctionnels de la Santé). Toutefois, une partie de ces données, sont des données constitutives de données plus générales sur les unités légales (Entreprises référencées par l'INSEE).

La sémantique des données de référence présente un enjeu considérable pour l'État. Un modèle sémantique robuste et générique devra être définie et entretenue, pour chaque ensemble de données de référence. C'est un facteur de succès indéniable pour tous les échanges de données au sein de l'État, mais aussi avec tous les partenaires, fournisseurs, et bien sur les usagers eux-mêmes (particuliers, entreprises, associations…).

3.4. Une métadonnée

Une méta-donnée est littéralement une donnée qui définit et/ou décrit une donnée, quel que soit son support, (titre, auteur, date de création par exemple). Ces métadonnées sont généralement définies sous la forme de dictionnaire ou de registre sur lequel le système d'information s'appuie pour « comprendre » des données utilisées par les différentes applications (définition sémantique d'une information de référence). C'est bien par la combinaison des données et de leurs métadonnées, qu'un acteur peut utiliser les données, qu'il soit un utilisateur (une être humain), ou une application informatique (un automate).

Ces métadonnées caractérisent la donnée sous-jacente : type de ressource, son format, qui l'a créé, quels sont les contributeurs, le sujet traité, la couverture sur laquelle s'applique la donnée (couverture géographique, temporelle, sectorielle…), des dates (liées au cycle de vie de la donnée : date de création, de mise à jour…).

Pour les ressources numériques, il existe des normes comme le registre *Dublin Core*[14] (ISO 15836) qui définissent un schéma de métadonnées génériques. Par ailleurs, le G8 a signé une charte[15] le 18/06/2013 rappelant les grands principes d'ouverture des données publiques, et en particulier proposant, en annexe de la charte, une définition commune des métadonnées, avec une correspondance entre la vision et le vocabulaire retenue par les 8 membres du G8. Il existe également pour des secteurs d'activités donnés comme la santé, ou la culture des normes de métadonnées spécifiques. Pour l'information géographique, la directive européenne INSPIRE normalise également les métadonnées associées.

À revoir la classification en deux : partir de l'OASIS

On distingue classiquement deux types de métadonnées :

- Métadonnées métier ; il s'agit de décrire l'organisation (classification) et les correspondances entre les objets métier vus par un utilisateur et les objets techniques correspondants. Ceci inclut les relations qui unissent les objets ainsi définis (liens père-fils d'une hiérarchie), ou des caractéristiques métiers propres au cycle de vie de l'objet métier.
- Métadonnées techniques ; il s'agit d'une construction permettant de documenter et de maîtriser (notamment en termes d'analyse d'impact, de conservation, de pérennité…) les structures manipulées dans les processus de chargement et de traitement des données (longueur d'un champ, définition d'un mapping …). Cela inclut les modèles de données.

Le travail de sémantique s'applique tout autant sur des données que leurs métadonnées.

3.5. Qualité des données

Il n'est pas inutile de le rappeler : les données jouent un rôle majeur dans l'exécution des processus et activités des acteurs au sein de l'État. Elles jouent également un rôle prépondérant sur le pilotage d'activités et de politiques publiques, puisque les décisions sont la plupart du temps fondées sur des données provenant d'outils décisionnels.

La qualité des données, et en particulier des données de référence, se définit comme l'aptitude de l'ensemble des caractéristiques de ces données à satisfaire des exigences internes (pilotage, décision, sécurité, efficacité, efficience…) et des exigences externes (réglementation…) à l'organisation. Il s'agit bien du « degré d'adéquation à l'usage que l'on en fait ».

Un des avantages majeurs du numérique est la communication et la réutilisation. Il n'est ainsi pas raisonnablement faisable de maîtriser toutes les utilisations possibles en aval. Plutôt que de chercher à maîtriser les exigences de tous les usages possibles, il est beaucoup plus pertinent de décrire, de mesurer et de communiquer sur le niveau de qualité atteint pour chaque donnée de référence, et, dans un processus d'amélioration continue, de progresser en communiquant sur le niveau visé (actions planifiées et en cours). Cela implique donc d'une part d'avoir défini des indicateurs de mesure de cette qualité et d'autre part d'avoir des outils pour mesurer et analyser ces indicateurs et agir en conséquence sur les données : analyse (recherche d'erreur, de doublon, d'amalgame…), nettoyage (correction, standardisation…), intégration (la réconciliation par exemple qui consiste à mettre en correspondances des données de différentes sources), enrichissement (suppression des doublons, fusion de données…), prise en compte des durées de conservation, suivi…

La qualité de données, et en particulier de données de référence peut se définir à travers un ensemble de critères limités :

- des critères intrinsèques aux données elles-mêmes
- des critères de services, liés à l'utilisation de ces données
- des critères de sécurité, globalement liés à l'ensemble du disponible de gestion de ces données

[14] http://dublincore.org/
[15] http://www.modernisation.gouv.fr/fileadmin/Mes_fichiers/pdf/Charte-G8-Ouverture-Donnees-Publiques-FR.pdf

Critères intrinsèques	Définitions
Unicité	Chaque entité du monde réel est représenté par un et un seul objet métier : absence de doublon et d'amalgame
Complétude	Les données sont complètes (instances et leurs caractéristiques) par rapport à la réalité et aux processus de collectes
Exactitude	Les données sont exactes et égales à la réalité qu'elles sont censées représenter. Ce critère englobe la notion de précision et de validité.
Conformité	Respect des contraintes liés aux données
Intégrité	Les relations entre les objets métiers sont cohérentes et présentes
Cohérence	Les caractéristiques des données sont cohérentes entres elles et avec les autres données considérées.

Figure 4 - Qualité des données : les critères intrinsèques

Critères Services	Définitions
Accessibilité	Facilité d'accès aux données (synchrone / asynchrone, unitaire / masse, push / pull, open API…).
Cohérence	Gestion dans le temps de la distribution des données (problématique de synchronisation et de resynchronisation).
Actualité	Rapport entre les données et le temps, ou degré de mise à jour des données. La fraîcheur des données doit tenir compte de la réalité organisationnelle et réglementaire.
Pertinence	Mesure de l'utilité d'un ensemble de données, ou de l'adéquation de ces données à des usages

Figure 5 - Qualité des données : les critères de services

Critères Sécurité	Définitions
Disponibilité	Aptitude du référentiel à remplir une fonction dans des niveaux de services définis Propriété d'une information d'être, à la demande, utilisable par une personne ou un système.
Intégrité	Les données doivent être celles que l'on s'attend à ce qu'elles soient, et ne doivent pas être altérées de façon fortuite ou volontaire. Propriété assurant qu'une information n'a pas été modifié ou détruit de façon non autorisée.
Confidentialité	Seules les personnes, les applications ou les processus autorisées ont accès aux informations qui leur sont destinées. Caractère réservé d'une information dont l'accès est limité aux seules personnes admises à la connaître pour les besoins de leurs missions, ou aux processus autorisés
Traçabilité (preuve ou imputabilité)	C'est la garantie de disposer des éléments qui apportent la preuve des traitements ou autres évènements relatifs aux informations considérées.
Lisibilité	le codage d'une information doit à tout moment être décodable. L'obsolescence des codages peut ainsi conduire à des recodages, dans l'objectif du maintient de cette lisibilité.

Figure 6 - Qualité des données : les critères de sécurité

Il existe dans la littérature, ou sur le web, une profusion de critères sur ce sujet. Le présent cadre en a retenu un sous-ensemble jugé pertinent pour couvrir les besoins en matière de qualité de données. Certains sont des incontournables mais il en existe de nombreux autres qui pourront s'avérer pertinents pour telle ou telle donnée dans un contexte particulier.

3.6. Un référentiel de données

Un **référentiel de données est un dispositif permettant la gestion mutualisée, et le pilotage dans le temps, de données de référence**, à savoir, l'actualisation et la mise à disposition de ces données. Un référentiel de donnée comprend aussi les éléments permettant de piloter et faire évoluer cette gestion de données de référence, en fonction des besoins des « clients » du référentiel, et en fonction des problèmes de qualité de données rencontrés.

Il se concrétise généralement par une architecture de données informatique (quelle que soit la technologie utilisée) qui contient les données de référence considérées. Mais attention il s'agit avant tout de considérer le référentiel fonctionnellement comme une brique mutualisée du système d'information, comme un ensemble d'éléments qui rend des services particuliers autour de la gestion de données de référence. Ces éléments ne sont pas tous nécessairement centralisés et uniques. Il existe des architectures de référentiel de données ou, par exemple, le stockage physique des données est distribué (sur des ressources physiques ou virtuelles différentes).

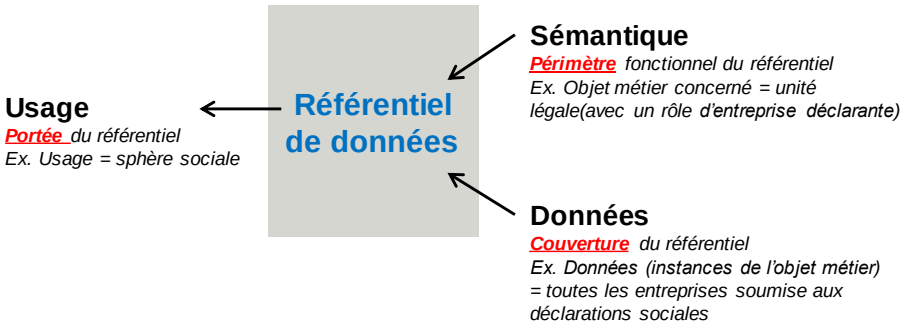

Figure 7 - Périmètre, couverture et portée d'un référentiel de données

Quand il est question de référentiel de données, il est nécessaire de préciser rapidement trois caractéristiques importantes : le périmètre, la couverture, et la portée du référentiel.

- Quelle est la sémantique des données de référence qu'il gère ? de quoi parle-t-on ? le terme « **périmètre** » à propos d'un référentiel, désigne le secteur fonctionnel (ou une partie) que le référentiel gère. Il est question ici du « contenant » (fonctionnellement et quelle que soit la technologie sous-jacente).
- Quels sont les données contenues dans le référentiel ? le terme « **couverture** » à propos d'un référentiel désigne l'ensemble de données, des instances, gérées par le référentiel. Il est bien question ici du contenu. Prenons l'exemple d'un référentiel dont le périmètre est le secteur fonctionnel « Agents » du POS. Il gère donc des objets métiers « agents ». La couverture précisera quelles sont les « instances » gérées du ou des objets métiers considérés : par exemple tous les agents (donc interne) du ministère de l'intérieur.
- Pour quel usage le référentiel est-il prévu ? le terme « **portée** » à propos d'un référentiel désigne globalement les organisations (ou l'ensemble des acteurs et de leurs SI) pour lesquelles le référentiel peut être utilisé et répondre aux besoins de fourniture de données de référence. Reprenons l'exemple précédent d'un référentiel dont le périmètre est l'agent, la couverture les agents du Ministère de l'intérieur, cela n'en fait pas pour autant un référentiel utilisable (une portée) pour tout le SI du ministère. Cela dépendra beaucoup du niveau de qualité des données, et donc des processus de collecte. La portée pourrait être réduite uniquement aux RH par exemple.

Comme il l'a été précisé précédemment, les principales caractéristiques d'une donnée de référence sont la sémantique, la qualité, la centralisation de l'identification, le partage et la réutilisation. Un référentiel de données comprend l'ensemble des dispositifs (notamment informatiques) permettant de gérer, de faciliter, et de piloter ces différents aspects. Une partie de ces dispositifs peuvent d'ailleurs être mutualisés entre plusieurs référentiels. Le modèle sémantique des données de référence, est l'élément central, et nécessairement unique. Il convient là aussi de bien distinguer la sémantique des données de référence, leur sens métier partagé, et la syntaxe de ces mêmes données dans les services d'échanges qui peut éventuellement prendre des formes particulières et différentes en fonction des usages.

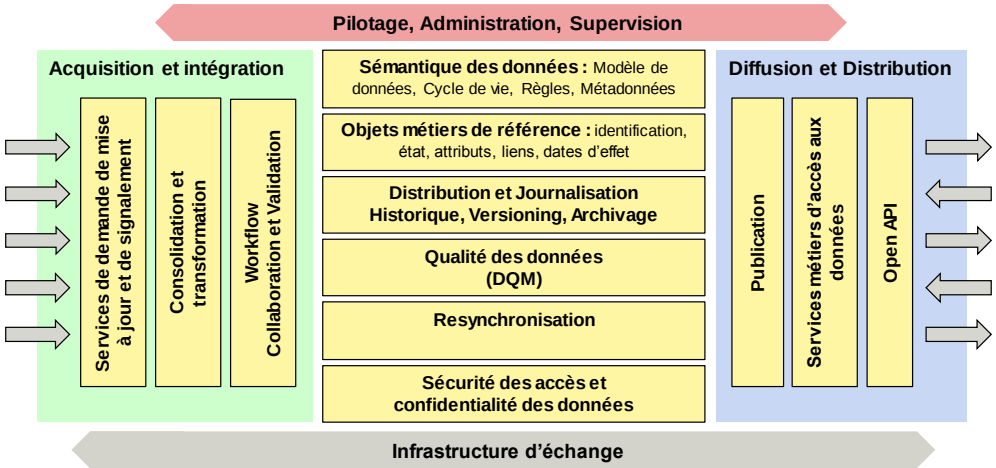

Figure 8 - Architecture générale d'un référentiel de données

Le chapitre suivant décrit l'architecture générale d'un référentiel, illustré dans la Figure 8 - Architecture générale d'un référentiel de données, ainsi que l'ensemble des règles de conception à appliquer par les équipes en charge de la construction et de la maintenance des référentiels de données de l'État.

Inversement, une base de données, contenant des données n'est pas nécessairement un « référentiel de données ». C'est avant tout la finalité (gérer et mettre à disposition des données de référence au bon niveau de qualité attendu), le type de données qu'elle contient, et les services qu'elle offre (notamment si ces services peuvent être considérés comme le point

de vérité, cf. ci-après), qui permet de qualifier si telle ou telle base de données est considérée, ou peut être considérée, comme un référentiel de données. Il arrive également qu'une application est dans ses usages considérée comme un référentiel de données, alors qu'elle n'intègre pas tous les dispositifs lui permettant d'assurer ce rôle : fiabilisation de la collecte en amont, services publics d'accès aux données, et services de resynchronisation. Le présent cadre doit permettre de faire évoluer de telles applications, en effectuant un diagnostic précis des forces et faiblesses de tels référentiels.

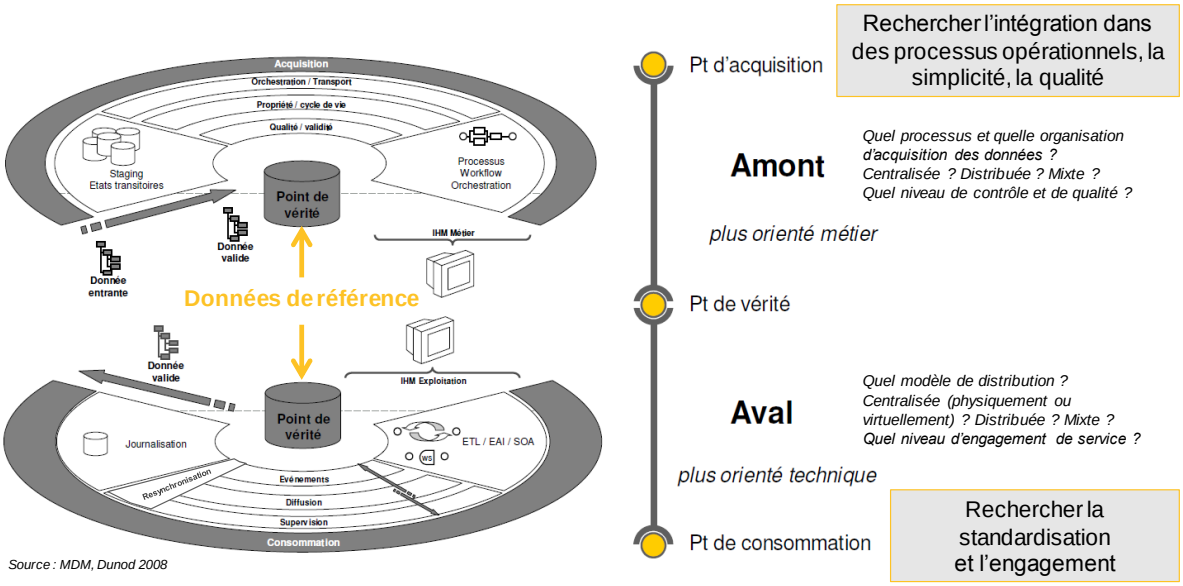

Figure 9 - Acquisition, point de vérité et consommation des données

En termes d'architecture la mise en place d'un référentiel consiste à créer un point de vérité au sein du système d'information qui doit garantir la validité des données de référence qu'il détient, c'est-à-dire, leur niveau de qualité intrinsèque (tel que définie précédemment). Il est nécessaire dans la mise en place d'une telle architecture de bien séparer les problématiques amont et aval à ce point de vérité, à savoir : l'acquisition des données, et la consommation (ou diffusion / distribution) des données. La Figure 9 - Acquisition, point de vérité et consommation des données, illustre les différentes problématiques à adresser lors de l'identification et la mise en place du point de vérité. C'est le lieu où les données de référence sont disponibles à tous, au juste niveau de qualité défini et communiqué. Le positionnement de ce point de vérité dans le système d'information peut être différent du ou des points d'acquisition des données, qui sont le ou les points d'entrée de chaque donnée dans le SI. Il est possible également que le lieu réel de consommation, le point de consommation, de la donnée diffère du point de vérité, généralement pour des raisons de sécurité ou de performance, mais également pour des raisons métiers. En effet, la notion de point de vérité est intimement lié aux états des données et donc à leur cycle de vie (cf. le paragraphe précédent sur la sémantique).

La définition de l'architecture d'un référentiel consiste donc précisément à définir ou positionner parmi les applications existantes, où devront se trouver les points d'acquisition, l'unique point de vérité et les éventuels différents points de consommation (il y en aura nécessairement au moins un). Les chapitres suivants définissent les règles à appliquer, et les architectures types (*design pattern*) possibles. Il est important de comprendre que la mise en place d'un référentiel est un sujet particulièrement complexe et nécessairement long. Car il s'agit bien de définir comment au sein de l'État, l'organisation et les processus métiers, et les outils informatiques se recentrent sur ce point de vérité de mise à disposition de données de référence nécessaires à tous.

La mise en place de référentiel de données ne consiste pas systématiquement à centraliser en un seul lieu toutes les données de référence utilisées par toutes les autorités administratives. Au niveau de l'ensemble de l'État, c'est tout simplement inenvisageable pour de nombreuses données de référence. Pour des raisons d'urbanisation et d'agilité, la centralisation à l'extrême est parfois contre-productive. Il est toutefois raisonnable de considérer que pour certaines données de référence, en particulier les nomenclatures externes, une architecture centralisée serait systématiquement préférable pour l'ensemble de l'État.

Dans tous les cas il est indispensable de considérer que, du point de vue du système d'information de l'État pris dans sa globalité, il est question aujourd'hui d'un ensemble de référentiel, interdépendants qu'il est nécessaire d'améliorer, d'optimiser, de simplifier, de pérenniser, de transformer. D'un point de vue de l'architecture logicielle et technique, il est par contre plus facile d'imaginer une homogénéisation des pratiques, une mutualisation des composants et des infrastructures.

Le chapitre 5 définit les différents types d'architecture recommandée.

Au-delà du dispositif informatique qu'est un référentiel de données, il est nécessaire de mettre en place une architecture d'ensemble, une organisation, des compétences, un pilotage pour suivre dans le temps la mise en place et l'évolution de référentiel de données. On parle alors de « gouvernance de données de référence » ou « master data management » (MDM). C'est précisément l'un des objectifs de la démarche permanente d'architecture d'entreprise, ou d'urbanisation.

3.7. Urbanisation du SI de l'État et référentiel de données

Le chapitre 2 a rappelé les objectifs de la démarche d'urbanisation du SI de l'État. Cette démarche vise à organiser la transformation du SI de l'État. Concernant les données de référence et les référentiels de données, elle vise à :

- Identifier et isoler fonctionnellement les données transverses (qui sont la plupart du temps des données de référence) dans une partie spécifique de la nomenclature de référence correspondante, le POS du SI de l'État. Il s'agit fonctionnellement de centraliser la gestion de données de référence : sémantique, gouvernance, et architecture. Et en particulier, il est nécessaire d'isoler le plus possible la gestion de données de référence des processus métiers qui les utilisent (principe de découplage).
- Impliquer les acteurs métiers à la définition sémantique des données de référence : en cherchant la vision partagée, si possible standardisée et sémantiquement riche.
- Evaluer le patrimoine actuel en matière de référentiels :
    - niveau de satisfaction des « clients » (des utilisateurs) des référentiels, maîtrise des processus d'alimentation et d'actualisation,
    - alignement de l'architecture applicative sur les règles définies dans le présent cadre,
    - niveau de qualité des données et de maîtrise de la qualité dans le temps,
    - niveaux d'engagement des acteurs.
- Proposer une cible de rationalisation, de simplification et d'optimisation métier, fonctionnelle et applicative, selon les règles définies dans le présent cadre. Il s'agit clairement de définir quelle architecture de distribution de données est la plus adaptée compte-tenu de l'existant, du contexte métier, de la cible à atteindre et des moyens disponibles.
- Définir une trajectoire de transformation vers la cible : intégrer dans les feuilles de route interministérielle et ministérielles les projets (ou sous-projets) d'évolutions de tel ou tel dispositif métier, fonctionnel ou applicatif concourant aux différents référentiels de données considérés.

Les référentiels de données constituent un élément clé du système d'information. Leurs transformations dans le temps doit donc être suivi et pilotée au plus haut niveau. L'architecture de référentiel de données et l'évolution de cette architecture est néanmoins un sujet particulièrement complexe. Car il s'agit très clairement de définir et de construire comment la sémantique des objets métiers de référence, manipulés par une grande partie des administrations, se matérialise et se distribue dans les outils informatiques et les échanges entre autorités administratives et comment ses dispositifs évoluent dans le temps par rapport aux évolutions légitimes des métiers de l'État et de leurs environnements.

La démarche d'urbanisation doit systématiquement capitaliser toute la connaissance sur les référentiels de données : sémantique, architecture, services, qualité des données… Cette connaissance en particulier sur les référentiels constitue un patrimoine clé.

En conclusion de ce chapitre et pour résumer, il est absolument nécessaire que tous les acteurs considèrent que les données constituent un actif pour l'État, et même un actif stratégique pour le pilotage et la transformation de politiques publiques. Dans de nombreux cas, et en particulier pour les données de référence, cet actif n'est pas spécifique à un métier ou une politique publique, mais transverse à plusieurs, et pour certains cas à tous les métiers de l'État. C'est le cas par exemple des données de référence sur les structures et les agents. Il est également tout aussi indispensable de comprendre que les données au sein de l'État n'ont d'existence qu'à travers le système d'information qui les collecte, les traite, les manipule, les archive. La gouvernance des données ne peut se faire sans lien avec la gouvernance du système d'information, plus vaste. Parler de données sans système d'information n'a aucun sens.

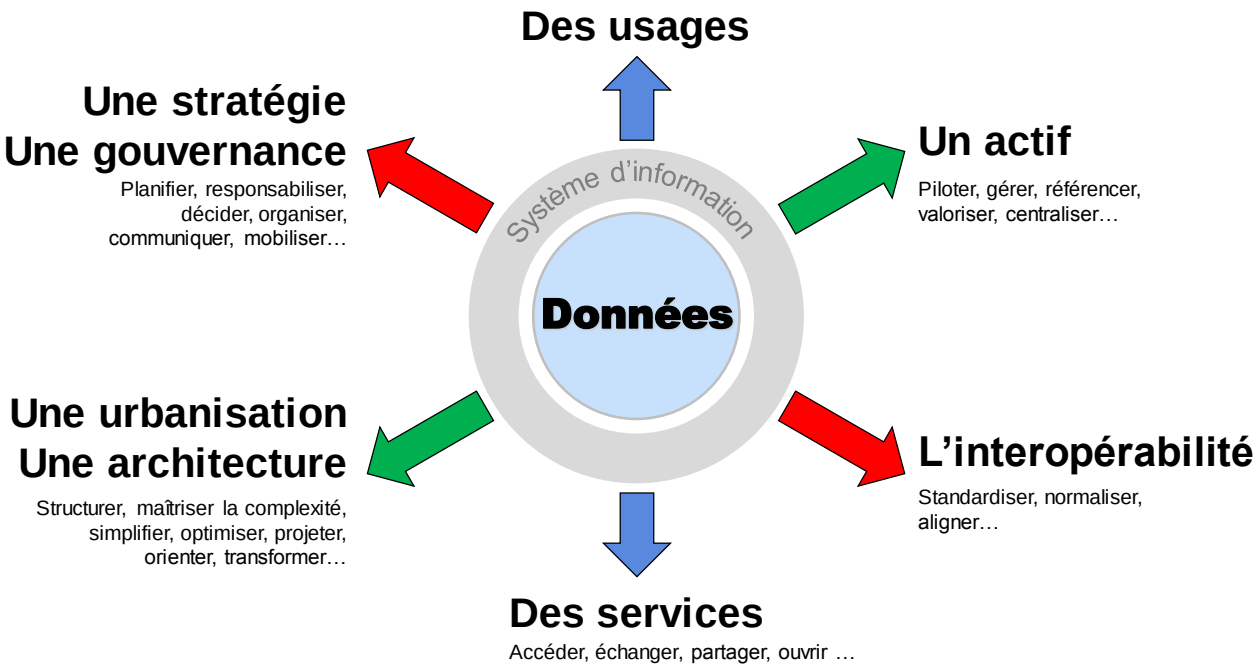

Figure 10 - Les six axes majeurs d'une gouvernance des données

De fait, le changement de paradigme devient nécessaire. L'approche doit être beaucoup plus globale, volontaire et structurée. La figure ci-dessus illustre les six axes majeurs qu'il est nécessaire de mettre en place, et qui sont développés dans ce document.

4. RÈGLES D'ARCHITECTURE DE DONNÉES DE RÉFÉRENCE

Le présent chapitre se compose de règles d'architecture à appliquer dans le cadre :

- D'évaluations des référentiels existants
- De projets de mise en place des nouveaux référentiels,
- De projets de transformation des référentiels existants
- De maintenances évolutives de référentiels existants

Les règles sont organisées selon les 5 vues du SI de l'État : Stratégie, Métier, Fonctionnelle, Applicative, Infrastructure. Elles sont structurées

4.1. Règles de niveau Stratégique

Les règles de niveau stratégique portent sur les actions de mise en place de référentiel, sur la gouvernance associée, et la performance de tels dispositifs.

Règle RS1	Toute donnée collectée, traitée, manipulée ou échangé au sein du SI de l'État et qui possède les cinq caractéristiques suivantes : • Grand nombre d'acteurs (utilisateurs, applications ou systèmes) internes ou externes utilisateurs et grandes fréquences d'utilisation • Qualité critique pour un grand nombre de processus • Sémantique partagée et relativement stable dans le temps • Durée de vie qui va au-delà des processus qui l'utilisent • L'efficacité et l'efficience globale des processus sont conditionnées par la facilité d'accès à la donnée …doit : • être considérée comme une donnée de référence, et donc un actif stratégique ; • être positionnée très clairement dans le POS du SI de l'État ; • être gérée dans un référentiel de donnée dont la portée doit être interministérielle dès lors que la donnée concerne au moins 2 ministères (en totalité ou en partie), et dont la gouvernance est opérationnelle. La montée en maturité de la gestion des données de référence est un objectif prioritaire dans la transformation du SI de l'État.

Il est capital d'identifier formellement les données de référence au sein de l'État et de les considérer comme des données clés. Le §3.3 décrit précisément les caractéristiques de ces données. Toutes ces données de référence doivent faire l'objet d'une attention toute particulière et être gérée au final dans un référentiel de données. La mise en place de cette gouvernance et de cet outillage peut être longue et coûteuse, mais elle doit dans tous les cas être identifiée, planifiée, budgétée et des ressources doivent lui être allouées.

La question de la portée du référentiel et de sa gouvernance est clairement au cœur de cette réflexion. Dès lors que les données de référence concernent au moins 2 ministères (pas nécessairement en totalité) différents, la portée interministérielle doit être **systématiquement** envisagée.

Au sein de l'État, il est probable que certaines données n'aient pas été formellement identifiées comme « données de référence », et ne sont donc pas gérées dans un référentiel, ou que certains référentiels n'aient pas la bonne portée nécessaire. Dans tous ces cas, il sera absolument nécessaire de planifier ce travail dans le cadre des contrats de progrès ministériels et interministériels.

Règle RS2	Le système d'information de l'État, et ses sous-systèmes, sont conçus de façon modulaire dans une approche globale : les référentiels de données sont isolés des outils métiers, des outils de pilotages ou de contrôles, et des outils de gestion des ressources (fonctions supports).

Les référentiels de données intègrent les principes d'urbanisation générale du SI de l'État. Ils sont fonctionnellement isolés dans le SI. Ils sont fonctionnellement isolés des domaines Opération, Ressources & Support, Echange & Relation, Pilotage & Contrôle, dans un domaine spécifique « Données transverses ».

Ce découplage est essentiel pour garantir, dans le temps, la cohérence des données, et donc des objets métiers, contenues dans ces référentiels. En effet leur cycle de vie va généralement bien au-delà des processus métiers qui les utilisent. Les fonctionnalités qui permettent de manipuler ces objets métiers doivent donc être isolées des fonctionnalités métiers courantes (opération, support, pilotage, relation…) pour éviter de surcharger ces dernières.

Règle RS3	Tout référentiel de données dispose d'un cadre de gouvernance à jour, accessible librement à tous, et aligné ou en cours d'alignement : • sur le présent corpus réglementaire interministériel (CCU[16], CCARD[17], RGI[18]), • sur la stratégie SI de l'État, • sur les besoins opérationnels des utilisateurs des données de référence gérées, • sur la réglementation métier correspondante en vigueur Une démarche d'amélioration continue est en place pour atteindre le niveau de qualité attendu par les utilisateurs du référentiel.

Tout référentiel de données doit disposer d'un cadre de gouvernance opérationnel. Le chapitre 6 est dédié à ce sujet. Au-delà des questions d'architecture (notamment fonctionnelle, et applicative), il est nécessaire de mettre en place « autour » de tout référentiel une gouvernance, c'est à dire le pilotage opérationnel qui permet de rapprocher les besoins des différents acteurs (source de données, consommateurs des données, réglementation, etc.) pour piloter la mise en place, l'évolution, voire la transformation, dans le temps du référentiel.

Ce cadre de gouvernance doit se matérialiser dans une comitologie et une documentation opérationnelle à jour comprenant :

- Identification claire du ou des responsables des données de référence,
- Sémantique des données (et des métadonnées),
- Processus général d'acquisition et de mise à jour des données : point d'acquisition, point de vérité,
- Engagement des acteurs MOA/MOE du référentiel, en particulier sur la qualité des données, sous forme de contrats de service,
- Architecture et catalogue des services.

La comitologie doit embarquer très clairement les décideurs métiers sur ces questions de pilotage et de stratégie de transformation. Cette comitologie peut s'imaginer sur trois niveaux :

- Le pilotage stratégique des référentiels : identification des responsables, la clarification des périmètres de responsabilités, la stratégie de transformation et donc les trajectoires associées, et bien sûr les travaux d'urbanisation.
- Le pilotage opérationnel centré sur la sémantique des données, leurs usages, et sur la conception et les évolutions des services des référentiels.
- Le pilotage technique centré sur la qualité des données et la maintenance des services des référentiels.

Règle RS4	Les dispositifs de mise à jour des données d'un référentiel doivent être ancrés dans des processus métiers opérationnels liés aux traitements du cycle de vie des objets métiers considérés. En fonction de la nature des données de référence, le juste compromis entre d'une part la réactivité des mises à jour (et des corrections) et d'autre part le contrôle à la source, doit être définie en toute transparence. Dans tous les cas, l'opérationnalité du référentiel doit être privilégié, à savoir : la réponse aux besoins opérationnels des utilisateurs des données de référence.

La qualité de conception et d'exploitation des processus de mise à jour des données d'un référentiel de données est un élément clé qui conditionne directement le niveau de satisfaction des utilisateurs du référentiel. Il est primordial d'avoir une vision précise du cycle de vie des données gérées dans le référentiel, mais pas uniquement une vision technique de ces changements d'états. Comme il a été décrit précédemment, c'est le travail de sémantique qui doit identifier et décrire les objets métiers gérés, leurs comportements, leurs différents états et changement d'états. La description sous forme de diagramme d'état de cette réalité métier (ou diagramme de machine à état), est un élément critique de la documentation

[16] CCU : Cadre Commun d'Urbanisation du système d'information de l'Etat.
[17] CCARD : Cadre Commun d'Architecture des Référentiels de Données, le présent document.
[18] RGI : Référentiel Général d'Interopérabilité.

sémantique. Il sera également bien sûr utile de la traduire techniquement pour les services de mise à jour et d'accès aux données.

Les processus d'acquisition et d'actualisation des données doivent impérativement être alignés sur le cycle de vie de ces objets métiers, et non l'inverse. Cela sous-entend de fait que sans cycle de vie clairement définie et décrit, il n'y aura pas de processus efficace d'acquisition et d'actualisation de données.

Dans certains cas, ce cycle de vie peut-être technique, comme par exemple l'élaboration d'une carte topographique.

Comme le sujet de ces processus est l'acquisition et l'actualisation de données massivement réutilisées au sein de l'état, avec une exigence en terme de qualité de données parfois très élevées, il est nécessaire de chercher l'automatisation au maximum de ces chaînes de valeurs, et de les encadrer avec des indicateurs permettant de piloter cette collecte.

Compte-tenu de la couverture et la portée des référentiels de données de l'État, il est également indispensable de privilégier une très forte réactivité dans la mise à jour et la correction des données de référence. La capacité des dispositifs de mise à jour doit le permettre. Ces dispositifs doivent intégrer en conséquence un mécanisme de signalement d'erreur ou de problème dans l'actualisation des données. La logique de *crowdsourcing*, à savoir, d'approvisionnement ou d'entretien par la multitude, est un mode de fonctionnement qui peut très clairement répondre à un de nombreux objectifs opérationnels des référentiels de données de l'État (par ex. sur la gestion des adresses).

C'est réellement un compromis qui doit être étudié, entre :

- D'une part la conformité des mises à jour du référentiel, avec un ensemble de contrôle à l'entrée, permettant de garantir la qualité (exactitude, complétude, actualité…) des données renseignées. Il est question de contrôle a priori.
- D'autre part la vitesse et la réactivité des mises à jour du référentiel, avec notamment, une boucle de retour (*feedback loop*) permettant de corriger rapidement toutes anomalies détectées par les utilisateurs. Il est question ici de contrôle a posteriori par l'ensemble des utilisateurs du référentiel.

Ce compromis, et les choix résultant, doit être très explicitement communiqué à tous les acteurs : producteurs, consommateurs et ré-utilisateurs des données.

Les 2 types de contrôles seront généralement à mettre en place simultanément, en privilégiant le premier ou le deuxième type de contrôle et donc de dispositif, en fonction de la nature et la criticité des données. Prenons un exemple sur le référentiel des données entreprises : il est clair que la création, et donc l'existence même d'une entreprise est un élément critique pour beaucoup de métiers et de processus au sein de l'État. Ce type d'évènement doit donc très clairement être encadré et la saisie des données contrôlées à l'entrée dans le SI de l'État (cas 1). Par contre les données de contact d'une entreprise (adresse postale et téléphonique par exemple) peuvent être considérées comme moins sensibles, et par ailleurs plus volatiles (ce n'est qu'un exemple, et il est clair que pour certains métiers ces informations peuvent devenir problématiques). Un contrôle a posteriori pour ce type de données serait suffisant (cas 2), surtout si la mise à jour est efficace et rapide. Une mise à jour possible (ou proposition de mise à jour) des moyens de contact d'une entreprise par l'ensemble des acteurs (en mode *crowdsourcing* donc) pourrait être un gage de qualité des données (notamment d'exactitude et d'actualité).

La règle RS3 précise les réglementations qui doivent encadrer la mise en place d'un référentiel, en particulier s'il s'agit de données à caractère privées ou sensibles. Mais il est évident que le responsable d'un référentiel a d'abord la charge de répondre aux besoins opérationnels de l'ensemble des utilisateurs. Il doit donc de fait être en mesure de proposer des évolutions de la réglementation pour lever toutes limitations ou contraintes qui ne permettraient pas de répondre efficacement aux besoins réels opérationnels des utilisateurs (et ré-utilisateurs). Prenons un exemple : le chiffre d'affaire d'une entreprise est une donnée utilisée par de nombreuses autorités administratives notamment dans l'attribution de droits et prestations. Or cette donnée est bien détenue en particulier par la DGFiP mais protégée par le secret fiscal, et détenue par l'INSEE mais protégée par le secret statistique. La levé encadrée de ces secrets permettrait réellement de simplifier la collecte de cette information pour de nombreuses autres administrations (protection sociale, travail, emploi, agriculture, environnement, etc.), et de simplifier la relation avec les entreprises elles-même.

Règle RS5	La qualité des données d'un référentiel est mesurée, suivie et publiée. Une démarche d'amélioration continue est en place pour atteindre le niveau de qualité attendu par les utilisateurs directs pour la performance de leurs processus métiers.

Il est illusoire de prétendre maîtriser toutes les utilisations possibles des données contenues dans un référentiel. En conséquence, prétendre maîtriser le niveau de qualité attendu par toutes ces réutilisations est tout simplement impossible. Il est par contre raisonnable, utile et nécessaire même, de communiquer sur le niveau de qualité atteint par un référentiel, et les actions d'améliorations en cours cherchant à améliorer tel ou tel critère de qualité.

Cela nécessite donc :

- la définition d'un certain nombre d'indicateurs de qualité ;
- la mise en place de composant de collecte de ces indicateurs ;

- la mise en place dans le cadre de la gouvernance d'un suivi de ces indicateurs (objectifs, actions correctives, améliorations, etc.) ;
- la mise en place d'un dispositif de publication (type tableau de bord).

C'est très clairement une condition de succès de l'utilisation d'un référentiel, et de la réutilisation des données et donc des services d'accès à ces données.

Règle RS6	Les référentiels respectant l'ensemble des règles du présent cadre pourront faire l'objet d'une Labellisation par la Direction Interministérielle des Systèmes d'Information et de Communication.

Les référentiels conformes au présent cadre, ou en cours de mise en conformité avec le présent cadre, pourront faire l'objet d'une labellisation par la DISIC. Les modalités et condition de labellisation ne sont pas définies à ce stade. Elle feront l'objet d'un document spécifique.

4.2. Règles de niveau Métier

Règle RM1	Unicité du point de vérité pour toutes données de référence.

L'objectif est d'identifier pour chaque donnée de référence l'unique point de vérité, c'est-à-dire, le lieu où les données de référence sont disponibles pour tous, au juste niveau de qualité défini et communiqué. Il s'agit d'identifier l'organisme responsable de la donnée, et le dispositif technique, donc le référentiel dans lequel le point de vérité de la donnée peut-être trouvé.

Pour une évidente efficacité du dispositif, le point de vérité est obligatoirement unique. Il faut distinguer la conception logique de ce point de vérité, nécessairement unique, et la réalité physique notamment au niveau des infrastructures de stockages, qui peuvent être multiples pour des questions de performances ou de sécurité par exemple.

Règle RM2	Les processus de mise à jour des données de référence sont décrits, publiés, entretenus et partagés. Ils identifient clairement tous les points d'acquisition des données et l'unique point de vérité. Les processus d'entretien des données sont alignés avec le cycle de vie métier des objets métiers.

Les points de collecte peuvent être multiples. La logique de consolidation doit être définie et intégrée dans des processus opérationnels décrits, formalisés et publiés et entretenus dans le temps.

Les points d'accès peuvent également être multiples, avec une traçabilité, une logique de distribution définie, et un modèle de synchronisation défini. Dans le cas de référentiel esclave d'un autre référentiel, la synchronisation des données doit être parfaitement définie, outillée et maîtrisée : c'est un élément clé pour ce type de référentiel.

Règle RM3	Les indicateurs de qualité sont identifiés dans les dispositifs d'acquisition et de distribution des données de référence. Ils sont mesurés et publiés.

Pour chaque référentiel de données, un ensemble d'indicateurs de qualité sont identifiés, mesurés et publiés au niveau des dispositifs d'acquisition, de stockage et de distribution des données. Un plan d'action est formalisé pour corriger les dysfonctionnements mesurés. L'objectif n'est pas de viser une qualité maximale de données, mais d'être transparent sur le niveau atteint : volume de données, complétude de la description des données, fraicheur, fréquence de mise à jour, intégrité, présence de doublon ou d'amalgame, etc.

Règle RM4	Chaque référentiel doit intégrer un dispositif d'alerte ou de signalement permettant à un utilisateur du référentiel de faire remonter au responsable du référentiel toutes anomalies sur les données détectées en aval (incomplétude, incohérence, doublon, amalgame, problèmes d'intégrité, etc.). Le processus de traitement des signalements doit être également décrit, à jour et publié. Il est recommandé également de rechercher son automatisation et donc son outillage.

La gestion des erreurs, ou de la mauvaise qualité des données est un sujet d'une complexité particulière. Il s'agit avant tout de disposer de processus outillés permettant de faire remonter aux équipes en charge de la gestion des données de référence de problèmes rencontrés ou détectés en amont (lors de la mise à jour par exemple) ou en aval de la chaîne (lors de la consultation des données). Il s'agit de disposer de services d'alertes appelables par tous les utilisateurs du référentiel pour signaler telle ou telle type d'erreur sur une donnée de référence. Ces erreurs doivent pouvoir être consolidées et publiées de manière transparente notamment dans le but de profiler (noter ou classifier) les données et faciliter ainsi la correction des erreurs ou en améliorer la qualité.

L'objectif est bien de créer un cercle vertueux entre la mise à disposition de données de qualité et la consommation de ces données. Le traitement de ces erreurs doit faire l'objet d'un processus formalisé en toute transparence.

4.3. Règles de niveau Fonctionnel

Règle RF1	La sémantique des données de référence est décrite, disponible, entretenue et partagée. Les objets métiers correspondant sont identifiés : leurs caractéristiques, leurs relations et leurs comportements (cycle de vie, états, événements) sont décrits et cette documentation est à jour par rapport à la réalité opérationnelle (processus et application).

La description précise des données de référence est un impératif dans tout dispositif de type « référentiel ». La première source d'erreur provient d'incompréhension ou de mauvaise interprétation de la sémantique des données. Ce travail de description de la sémantique doit comprendre :

- un volet statique : des définitions et des relations entre les concepts manipulés
- un volet dynamique : des états et des événements entre états, permettant de décrire le cycle de vie des données
- un volet règle de gestion : qui complète les deux précédents volets, et décrit des caractéristiques ou des comportements métiers particuliers. Il s'agit généralement de règles liées à la législation en vigueur.

Règle RF2	Rechercher une sémantique riche et un haut niveau d'abstraction. Formaliser cette sémantique à l'aide d'une modélisation selon la notation UML 2.4.1 (ISO)

Le travail de modélisation ne doit pas être considéré comme un luxe, ou une option facultative, mais il est au cœur de la maîtrise des données de référence : la représentation formelle et graphique du sens des données est indispensable. Ce travail de modélisation est un outil efficace pour identifier plus facilement les abstractions réalisables. Cette montée en abstraction est nécessaire pour rendre le modèle plus robuste, agile et flexible dans le temps. Elle permet également de simplifier le modèle en identifiant les caractéristiques ou comportements communs ou similaires. Enfin, ce travail facilite la communication, l'échange et l'appropriation de la sémantique par l'ensemble des acteurs concernés.

Règle RF3	La structuration sémantique des objets métiers est standardisée et basée sur le RGI.

Le RGI s'impose à tous les acteurs publics (administrations d'État, sphères protection sociale et santé, et collectivités territoriales), et donc à la fois aux acteurs en charge de la mise en place de référentiel, qu'aux utilisateurs de ces référentiels. Il définit les normes et standards applicable en matière d'interopérabilité (selon 4 volets : organisationnel, sémantique, syntaxique, technique).

Le RGI peut donc de fait être considéré comme le réceptacle des modèles sémantiques des données de référence utilisées au sein du SI de l'État. Ce qui signifie que les travaux de modélisation sémantique, réalisés dans le cadre de la mise en place d'un référentiel de données au sein du SI de l'État, ont vocation à enrichir le RGI.

Règle RF4	Séparer les données d'identités (ou d'identification métier), des identifiants des données de référence. Pour un objet métier : utiliser un identifiant de type URI[19] : aisément partageable, non ambigus, non signifiant (donc ne contenant pas de données élémentaires à caractères personnels ou potentiellement confidentielles), non modifiables, non-réaffectable, non supprimable et persistant.

La notion d'identifiant, donnée permettant d'identifier avec certitude un objet métier (une personne par exemple, ou une entreprise), doit très clairement être dissociée des données d'identités de l'objet métier.

Les « données d'identités » sont des informations permettant d'identifier une occurrence d'un objet métier, ce ne sont pas des données techniques. Le « nom », « prénom », « date et lieu de naissance » sont très souvent considérés comme des données d'identités d'une personne : ensemble elles permettent d'identifier avec une certitude proche de 100%, une et une seule personne. Par contre, elles ne peuvent et ne doivent pas être utilisées comme identifiants.

Cette notion d'identifiant est très souvent considérée comme une information d'immatriculation voire une information technique, mais avec l'usage de plus en plus répandue du numérique, cette information finie par être utilisée très fréquemment par les agents ou les usagers (ex. le n° de SIREN d'une entreprise). Il ne faut donc pas nécessairement l'exclure.

[19] Un URI, de l'anglais *Uniform Resource Identifier*, soit littéralement identifiant uniforme de ressource, est une courte chaîne de caractères identifiant une ressource sur un réseau (par exemple une ressource Web) physique ou abstraite, et dont la syntaxe respecte une norme d'Internet mise en place pour le *World Wide Web* (voir RFC 3986). Dans tous les travaux actuels autour des référentiels de données, du web sémantique, de l'administration des données, du master data management, il y a une très forte convergence sur l'utilisation de ce type d'identifiant.

La mise en place d'identifiant, ou de clé, permettant de retrouver avec certitude un objet métier, doit répondre à des exigences précises :

- cet identifiant doit être facilement partageable (dans un format interopérable) ;
- il doit être non ambigu ;
- il doit être non signifiant, c'est-à-dire ne contenant pas de données métiers ou techniques susceptible d'évoluer dans le temps, ne contenant pas de données à caractères personnels ou confidentielles ;
- il doit être non modifiable : une fois défini et attribué, il ne doit plus changer ;
- il ne doit pas être réaffecté à un autre objet métier, même si le précédent objet n'a plus lieu d'être (quelle qu'en soit la raison) ;
- Il ne doit pas être supprimable, même si l'objet n'a plus lieux d'être. (ex. L'identifiant d'une entreprise qui fait faillite, ne doit pas être supprimé, il est conservé jusqu'à la date légale de conservation, et même probablement au-delà dans cet exemple).
- Il doit être persistant : c'est-à-dire qu'il doit être réellement stocké, conservé et archivé dans le temps.

Dans le cas d'objets de type Acteur (Personne, Organisation, Unité légale) par exemple, mais cela peut être le cas pour d'autres données, il ne faut pas exclure la possibilité de gérer plusieurs identifiants, pour un même objet. Cet état de fait résulterait de plusieurs processus différents d'identification menés sans concertation. Un travail d'alignement devrait donc être conduit pour simplifier voire unifier cette identification.

Règle RF5	Toute donnée de référence, et donc tout objet métier ou toute relation entre objet métier, est positionné dans un secteur fonctionnel unique de la nomenclature de référence fonctionnelle du SI de l'État (NRF ou POS du SI de l'État). Ce positionnement des objets métiers et de leurs relations conditionne directement les dépendances entre données de référence, et donc les dépendances et interfaces entre les référentiels sous-jacents. C'est un acte clé en matière d'architecture d'ensemble du SI de l'État.

Il s'agit d'une règle fondamentale d'urbanisation. Chaque objet métier n'est positionné que dans un et un seul secteur fonctionnel du POS du SI de l'État. Il est également important de bien positionner les relations que ces objets métiers ont entre eux.

Par exemple, une Personne dispose d'une adresse postale de contact : il y a donc bien, en simplifiant le sujet, 2 objets métiers « Personne » et « Adresse » qui sont liés entre eux. Ces 2 objets seront très certainement positionnés dans des secteurs fonctionnels différents. Il s'agit également d'identifier le secteur qui est porteur de la relation entre ces 2 objets métiers. Cette question pose également le périmètre de responsabilité des acteurs, et le périmètre des référentiels.

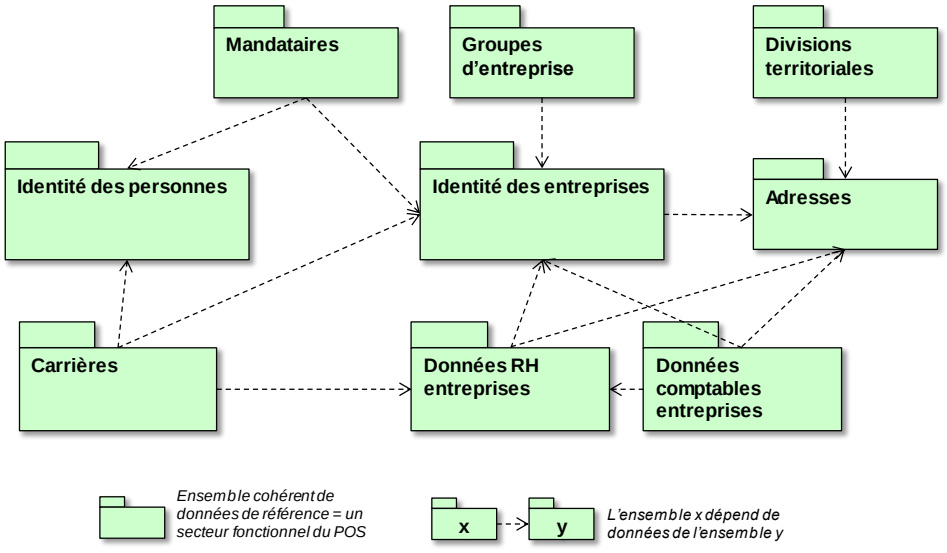

Figure 11 - Exemple de dépendances entre données de référence de la sphère sociale

La figure précédente illustre les dépendances entre ensemble de données de référence. Ce schéma n'est qu'un exemple, et ne prétend pas être complet et juste. Il met en évidence par exemple que :

- les données de référence sur les carrières (salaires et emplois des salariés) dépendent de données de référence sur les personnes, sur les entreprises et sur les données RH des entreprises.
- les données de référence comptables des entreprises, dépendent bien sûr de l'identité des entreprises, mais aussi d'adresse, et de données RH (les salaires).

- Les données de référence sur les divisions territoriales dépendent de données sur l'organisation du territoire et de données adresses.

Ce schéma ne fait pas apparaître les données de référence de type nomenclature, qui sont utilisées potentiellement dans chacun de ces ensembles pour qualifier ou structurer ces données.

Ce travail d'identification et de définition des ensembles (secteur fonctionnel + sémantique), et des relations entre ensemble est fondamental. Il conditionne directement l'architecture des référentiels de données. A ce niveau de l'architecture fonctionnelle, il est rappelé que chaque objet métier et chaque association n'est positionné qu'une et une seule fois dans un secteur fonctionnel.

Dans la figure 11, trois ensembles de données de références qui ont une particularité (identité des personnes, identité des entreprises, adresses). Ils conditionnent les autres ensembles. C'est-à-dire que les autres ensembles de données dépendent de ces trois-là. Il est donc important de sécuriser et de fiabiliser ce type d'ensemble de données de références. On parle alors de données de références maîtres ou socles, fonctionnellement, et donc de référentiels maîtres ou référentiels socles, applicativement.

Règle RF6	Toutes données de référence est sous la responsabilité du Responsable du Secteur Fonctionnel identifié.

Chaque secteur fonctionnel, et plus généralement, chaque zone, est sous la responsabilité et la gouvernance d'un Responsable de Zone Fonctionnel. Il est primordial que la responsabilité soit calée au niveau de chaque zone, mais il n'est pas exclu de la définir à un niveau plus fin (quartier ou bloc).

Cf. la version 1.0 du Guide du RZF

Règle RF7	Les fonctionnalités d'un référentiel de données (consultation et mises à jour des données de référence, identification, signalement d'anomalie, synchronisation et resynchronisation, consultation des indicateurs qualités, etc.) sont définies, documentées, entretenues et communiqués aux utilisateurs (consommateurs) de ces données. La cohérence des données doit être garantie par ces fonctionnalités et uniquement par elles. Il est exclus de pouvoir accéder aux données d'un référentiel par un autre moyen que des services publics qui implémentent ces fonctionnalités.

Toutes les fonctionnalités de mise à jour des données de référence doivent être conçues sans état. C'est-à-dire que l'exécution de ces fonctionnalités ne doit pas dépendre de résultats d'autres traitements (et donc d'autres fonctionnalités).

La qualité des données de référence, c'est-à-dire « le degré d'adéquation à l'usage que l'on en fait » est une exigence absolue. Des données de référence dont la qualité est mauvaise perdent très vite leur statut de « données de référence », et le référentiel qui les héberge, son titre de « référentiel ».

Il est donc primordial de mettre sous contrôle toutes les modifications sur ces données, aussi minimes soient-elles. C'est un des principes fondamentaux en matière d'architecture : l'encapsulation. Les accès aux données ne peuvent se faire qu'à travers un ensemble « d'interfaces », de services qui implémentent des fonctionnalités, et qui manipulent donc ces données de référence. Ce sont ces services qui garantissent l'intégrité de ces données, et les protègent en quelque sorte. Il est donc totalement exclu de pouvoir accéder aux données de référence par un autre moyen que par ceux prévus et testés (par exemple, un accès direct aux tables de la base de données d'un référentiel est à exclure totalement, même en lecture).

Note : Le présent document ne décrit pas les règles applicables à la conception des services. Le lecteur se référera en priorité aux principes d'architectures orientées services (SOA). Ce style n'est pas nécessairement l'unique matière de concevoir un référentiel (en particulier pour des référentiels dont l'échelle de temps est grande, avec de faibles échanges, ou l'utilisation d'outils de type ETL par exemple peut se concevoir).

4.4. Règles de niveau Application

Les règles suivantes s'appliquent aux les éléments de la vue applicative du système d'information, à savoir sur la conception et la maintenance du logiciel au sens large et de son architecture.

Règle RA1	Un référentiel de données porte sur des objets métiers d'un seul secteur fonctionnel de la Nomenclature de Référence Fonctionnelle (NRF ou POS du SI de l'État) : il encapsule ses données.

Il est primordial de limiter le périmètre fonctionnel d'un référentiel pour éviter toutes dérives de performances, de qualité, de pilotage du dispositif ou même d'usage. L'architecture fonctionnelle en amont, structurée notamment par le POS du SI de l'État, impose un premier niveau de découpage, et donc de fait un découplage des différents référentiels de

données. Il est primordial de s'y tenir sous peine de complexité croissante et donc de coût croissant très rapidement en fonction des objets métiers ajoutés.

Un référentiel doit par ailleurs rester vivant, dans le sens où, il doit être suffisamment agile face aux évolutions de son environnement (sur la réglementation par exemple) et des besoins de ces utilisateurs. La tentation d'élargir son périmètre fonctionnel, dans ces évolutions, peut-être grande. Il est fondamental de gérer de façon continue et précisément le périmètre du référentiel. Le Responsable de la Zone Fonctionnelle (RZF) a de fait un rôle primordial dans ce travail.

Cette modularité s'impose donc dans le temps : il est préférable d'avoir plusieurs référentiels dont le périmètre de chacun est clairement maîtrisé, dont la sémantique est parfaitement définie (ce qui permettra ainsi une plus grande robustesse, et flexibilité), plutôt qu'un seul dont le périmètre est hors de tout contrôle, et dont le coût d'entretien ne fera que croître.

Cette modularité fonctionnelle, n'impose pas de dupliquer tous les composants logiciels (vue applicative) et les infrastructures d'exécutions, de stockage et d'échanges (vue infrastructure). Ils peuvent bien au contraire être réutilisés ou mutualisés.

Règle RA2	Les données de référence encapsulées dans un référentiel ne sont accessibles que par des services documentés et catalogués, disposant de SLA partagés et alignés sur les besoins des consommateurs des données. Les services réalisent les fonctionnalités définies (RF7). Le mode d'accès synchrone est recommandé ; le mode asynchrone n'est pas interdit, mais doit être limité et encadré.

Le seul et unique moyen d'accéder aux données d'un référentiel est de passer par les services définis et documentés à cet effet. Ces services implémentent les fonctionnalités définies dans l'architecture fonctionnelle du référentiel (et uniquement celles-ci).

Tout accès par un autre moyen, non définie et sous contrôle de l'équipe en charge de la conception et de l'entretien du référentiel doit être rigoureusement interdit. La notion de service est à prendre au sens large indépendamment du style d'architecture logique (cela ne comprend donc pas que les « services web ») : par exemple, un dispositif d'extraction asynchrone de l'ensemble des données du référentiel, pour des besoins de resynchronisation, est également un service, s'il est associé à des engagements (disponibilité, qualité, maintenance…).

Les niveaux d'engagement de ces services doivent être définis dans des contrats de services (SLA), partagés et alignés sur les besoins des principaux consommateurs des données.

Dans le cas où les données de référence sont également des données publiques[20] (par exemple les données sur les entreprises du référentiel SIRENE), il est nécessaire de prévoir un ensemble d'API[21] (d'interfaces d'accès) ouverte – ou *open API*, à savoir un ensemble de services publics et donc ouverts, interopérables et accessibles depuis internet. L'objectif étant de pouvoir industrialiser la mise à disposition de ces données publiques, et donc d'interconnecter directement les logiciels des ré-utilisateurs de ces données avec le référentiel. Comme ces ré-utilisateurs peuvent tout aussi être dans la sphère publique que la sphère privée, il est nécessaire, plus simple, et finalement plus efficient de prévoir ces services d'accès sous forme d'*open API*.

Il est recommandé également de prévoir des services dédiés au signalement d'erreur. Ces services devront également être conçus de manière ouverte (open API) quand ils portent sur des données publiques.

Dans le domaine des données géographiques, la directive européenne Inspire[22] exige la mise en œuvre de services de recherche, de consultation et de téléchargement des données. Le téléchargement des données est généralement réalisé de manière asynchrone en raison du volume des données et de la capacité des réseaux.

Règle RA3	Toute application métier qui a besoin d'une donnée de référence doit faire appel au référentiel les encapsulant et utiliser les services disponibles. Il s'agit de réutiliser systématiquement les données de référence, et les référentiels labellisés.

Cette règle complète la précédente (RA2). Les référentiels sont des composants fondamentaux et structurants du système d'information de l'État. La mise en place de ces composants et leur entretien peut présenter dans certains cas des coûts importants. L'intérêt et la valeur apportée par de tels dispositifs n'est maximale que :

- si les ressaisies de ces données sont interdites (elles ne sont saisies qu'une seule fois dans la chaîne de collecte amont au référentiel),

[20] Se référer aux textes en vigueur pour la définition de « données publiques », notamment la circulaire n°5677/SG du 17 septembre 2013 du Premier Ministre et le « vademecum » associé.

[21] Une API ou *Application Programming Interface* est un ensemble normalisé de services qui sert de façade, d'interface, par laquelle un logiciel offre des services à d'autres logiciels. Ces API permettent ainsi des interconnections directes de logiciels à logiciels, et plus globalement de systèmes à systèmes. Elles peuvent prendre la forme de *services web* ou de *service REST* par exemple.

[22] La directive européenne 2007/2/CE du 14 mars 2007, dite directive Inspire, vise à établir une infrastructure d'information géographique dans la Communauté européenne pour favoriser la protection de l'environnement.
http://www.developpement-durable.gouv.fr/La-directive-europeenne-Inspire-de.html

- si ces référentiels sont réellement utilisés par toutes les applications du SI qui ont besoin de manipuler ces données de référence.

Règle RA4	La contextualisation des données de référence, et donc la conservation ou la copie locale dans une application métier, ou un autre référentiel, doit absolument être limitée et considérée comme une exception. Elle n'est possible qu'avec un encadrement strict et défini, par le Responsable du Secteur Fonctionnel (RZF) correspondant aux données de référence. La recopie de copie de données de référence est strictement interdite.

Cette règle complète les deux règles précédentes (RA2 et RA3). Un référentiel expose ses données à travers de services (RA2). Les applications utilisatrices des données de référence ne doivent en aucun cas les ressaisir et n'accèdent à ces données qu'à travers les services offerts par le référentiel (RA3).

Toutefois, les données de référence pouvant être dépendantes les unes des autres (cf. la règle RF5), il s'agit ici d'encadrer la conséquence de ces dépendances au niveau applicatif, en particulier dans la duplication ou non, et donc la synchronisation, des données.

Il est nécessaire de limiter au maximum et d'encadrer très précisément la (re)copie et la contextualisation des données de référence (en dehors des identifiants bien sûr), c'est-à-dire leur conservation et leur enrichissement par un acteur, une application ou système local.

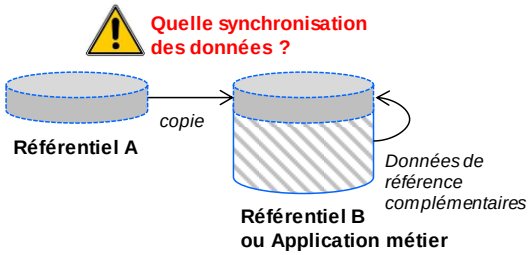

Figure 12 - Synchronisation de données de référence : une pratique à encadrer

Cet enrichissement signifie généralement une conservation (donc un stockage) locale de ces données pour en ajouter d'autres. La question de la synchronisation des données entre les deux dispositifs se pose clairement, et engendre une complexité de gestion forte.

Le RZF du secteur fonctionnel doit répertorier, encadrer, voire limiter le plus possible ces pratiques.

Au-delà de cette contextualisation, il est nécessaire également de préciser un point important. Les questions d'architecture de distribution et de synchronisation de données, font très certainement partie des sujets les plus complexes à traiter en informatique. Plus le système considéré est de taille importante, et à l'échelle de l'État son périmètre même n'est pas clairement défini, plus ce sujet de distribution et de synchronisation prend une ampleur, une combinatoire qui peut engendrer des coûts de mise sous contrôle considérable.

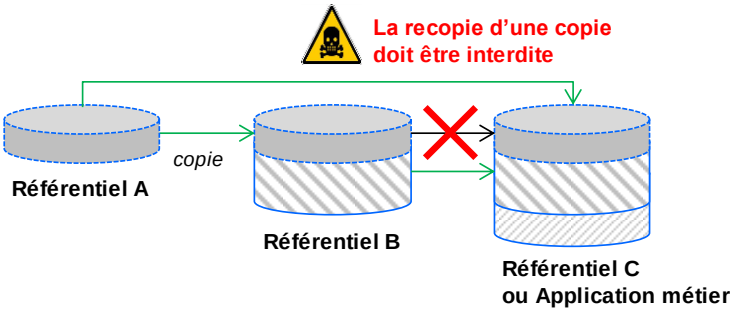

Figure 13 - La recopie d'une copie de données de référence est interdite

Le présent cadre assume totalement donc la règle stricte suivante pour limiter de manière cette complexité : **La recopie de données de référence d'un référentiel qui a lui-même été alimenté par un autre référentiel doit être strictement interdite**. NdA : cela ne concerne bien évidemment pas les identifiants des objets métiers, qui peuvent, et doivent même être recopiés.

Règle RA5	Un référentiel de données doit garantir la continuité des services qu'il offre.

Il l'a été dit à plusieurs reprise dans ce document, un référentiel de données est un élément pivot ou focal d'un système d'information. De fait, il doit être un composant relativement stable de l'ensemble du système, sous peine de fragiliser l'ensemble. Toutefois, il est illusoire d'imaginer qu'il ne peut ou ne doit pas évoluer. Les changements réglementaires, l'évolution de l'environnement, ou des réforment profondes imposent aux référentiels d'évoluer.

Le travail sur la sémantique, notamment la montée en abstraction, doit permettre de limiter l'impact de ces changements. Les règles fonctionnelles globalement permettront de contenir ces impacts dans le temps.

Quand l'évolution est finalement obligatoire, et que les services d'accès aux données doivent faire l'objet d'adaptation ou de transformation, il est primordial d'assurer la continuité de services dans la durée. Les services seront donc versionnés, avec un fonctionnement en parallèle de plusieurs versions. Il est toutefois recommandé de ne maintenir qu'au maximum deux versions en production. Au-delà la complexité de maintenance, de migration, croit de manière beaucoup trop importante.

En corollaire, cela signifie que toutes les applications métiers utilisatrices d'un référentiel doivent également faire le nécessaire pour suivre les évolutions des services de ce référentiel, pour ne pas pénaliser l'ensemble du dispositif (et éviter ainsi un effet « maillon faible »).

Règle RA6	Toute action sur les données d'un référentiel doit être tracée et journalisée. Les données sont historisées. Les instances des objets métiers sont versionnées en fonction de leur cycle de vie. Les données ne sont pas détruites, mais marquées comme non active. La politique de conservation et d'archivage des données est définie.

Un référentiel de données étant un point de passage systématique pour de nombreux processus métiers, il doit comprendre de fait un certain nombre de dispositifs, qui seront à concevoir en fonction notamment de l'analyse de risque :

- la sécurisation des accès aux données et des données : gestion de l'authentification et des autorisations d'accès pour les utilisateurs (personnes, applications, systèmes) ; gestion de la confidentialité des données ; gestion de scellement et de certification de données (apporter l'assurance que la donnée est conforme au référentiel).
- la traçabilité des accès et des modifications des données : le référentiel doit inclure des pistes d'audit permettant d'analyser les accès aux données, les dysfonctionnements ou pour répondre aux besoins de contrôle.
- les données elle-même doivent être versionnées en fonction du cycle de vie de l'objet métier correspondant. Il pourra être utile de figer certains états de l'objet métier dans le temps, et d'historiser donc les modifications sur les données (ce qui permet d'obtenir des photographies à date des données de référence). En particulier, les suppressions devront être gérées logiquement. Une donnée de référence ne doit pas être détruite (supprimées physiquement), mais marquées comme non active (suppression logique). Ce qui implique la mise en place de « date d'effet ». Il convient donc de mettre soit une gestion de « photo » d'objets métiers, soit une gestion de dates d'effets, voire une combinaison des deux. Dans tous les cas, il convient d'être très vigilant sur la documentation, la communication et la compréhension de ces dispositifs auprès des utilisateurs du référentiel.
- une politique de conservation et d'archivage des données doit être définie.

Attention, un référentiel n'est toutefois pas un dispositif d'analyse et d'aide au pilotage (décisionnel). Il n'a pas vocation à conserver toutes les images cohérentes des objets métiers gérées. D'autant que la définition de ces images dépend des consommateurs. Ces besoins seront traités par un autre dispositif que le référentiel (cf. le cadre d'architecture commun qui sera réalisé sur le décisionnel).

Règle RA7	Tout utilisateur (ex. une application métier) du référentiel doit pouvoir effectuer une (re)synchronisation totale ou partielle des données de référence qu'il stocke localement.

Le fait d'autoriser les éventuelles copies de données de référence dans des applications métiers ou d'autres référentiels, c'est-à-dire d'autoriser une forme de distribution des données de référence, pose clairement la question du maintien de la cohérence dans le temps de ces données avec le référentiel d'origine (cf. RA4).

Les raisons de ces incohérences ou désynchronisations sont multiples : des erreurs des saisies mal diagnostiquées, des erreurs dans l'exécution des processus, des erreurs informatiques (des bugs non détectés, un mauvais filtrage, un crash, une mauvaise restauration...). En théorie, les contrôles doivent être réalisés systématiquement le plus en amont possible pour éviter ces différents cas d'erreurs. Dans la pratique, et en particulier à l'échelle de l'État, **il est vital de considérer que ces contrôles ne seront jamais suffisants, et que des mécanismes de resynchronisation sont donc indispensables**.

Ce sujet doit être traité dès la conception d'un référentiel. Il doit se traduire par la définition et la mise en place d'un dispositif permettant à tout utilisateur d'effectuer la re-synchronisation volontaire des données copiées.

Le but de cette resynchronisation volontaire est de maintenir la cohérence des données contenues dans des localisations secondaires (application métier ou référentiel esclave) par rapport à un référentiel « source ». Il s'agit donc de vérifier la cohérence entre 2 ensembles de données, et de « nettoyer » les données copies en conséquence : c'est-à-dire de vérifier et de refléter sur la copie toutes les modifications intervenues sur la source, pour un même périmètre fonctionnel et une même couverture de données.

Ce besoin de resynchronisation peut être :

- ponctuel, par exemple dans le cas de détection d'erreur, ou encore lors de montée de version (du référentiel ou de l'application métier),
- mais aussi permanent ou périodique,

En conséquence le modèle de cohérence sous-jacent doit être souple, et pas nécessairement « strict » avec des mécanismes lourds. Cette remise à niveau, ou nettoyage des données, peut nécessiter des arbitrages et donc l'implication du propriétaire des données. Même si la recherche de l'automatisation dans ces mécanismes est recherchée, un tel dispositif peut être en partie manuel.

Le présent cadre ne définit pas les méthodes ou solutions à mettre en place pour de tels dispositifs. La complexité de tels dispositifs est bien réelle, d'où la nécessité de les définir, les concevoir, et les mettre en place le plus tôt possible.

Dans le domaine des données géographiques, les outils actuels ne proposent pas de mécanismes de resynchronisation.

En conclusion, un référentiel de données doit donc intégrer des services :

- De vérification (présence et qualité des données copiées par rapport à la source, mais également détection de données présente dans la source et absente de la copie).
- De rapprochement, ou d'identification (quand il y a une erreur ou une absence d'identifiant).
- De réplication ou d'asservissement des données (qui peuvent prendre différentes formes).

4.5. Règles de niveau Application et Infrastructure

Règle RI1	Respect strict des exigences du RGI sur les volets techniques, syntaxique

En soit cette règle est inutile puisque le Référentiel Général d'Interopérabilité s'applique à tous les acteurs. Elle a été délibérément ajoutée dans ce cadre d'architecture pour insister sur la nécessité absolue de respecter les exigences techniques et syntaxiques du RGI dans le cadre de la construction et l'entretien de référentiel de données.

Un référentiel de données est un élément pivot fondateur dans un système d'information. Il est donc primordial qu'il soit conforme en tout point à la dernière version en vigueur du RGI.

Règle RI2	Banalisation des infrastructures logicielles et mutualisation possible des composants des référentiels de données de l'État… vers la mise en place d'une infrastructure MDM commune.

Les référentiels sont des composants clé d'un SI. Il est donc indispensable de sécuriser au maximum l'infrastructure d'exécution, de stockage et de communication qui les supportent. La mutualisation de ces infrastructures logicielles et physiques doit être systématiquement recherchée, car les investissements en la matière peuvent être conséquents d'une part, et que d'autre part, il n'y a pas de différence significative techniquement entre deux référentiels.

Les mutualisations doivent être envisagées jusqu'au niveau de l'hébergement. Certains référentiels nécessiteront une très haute disponibilité et tolérance aux pannes, il est donc nécessaire de mutualiser ces coûts et minimiser les risques.

Les nomenclatures de référence fonctionnelle, applicative et infrastructure (NRF, NRA et NRI) permettent de cadrer ces mutualisations. La mutualisation des composants techniques ne signifie absolument pas que tous les référentiels doivent être fusionnés en un seul. Il est important de conserver la modularité et l'agilité nécessaire en séparant fonctionnellement les différents référentiels : les règles RF5 et RA1 sont applicables.

5. ARCHITECTURES TYPES D'UN RÉFÉRENTIEL DE DONNÉES

Le présent chapitre décrit un premier ensemble d'architecture type, appelé *design pattern*, recommandés dans la mise en place de référentiel de données. Il s'agit uniquement de modèle de conception logique. Chacune ces architectures peut ensuite d'un point de vue d'une architecture d'exécution se décliner de plusieurs manières, en fonction des exigences notamment de performance, de disponibilité, de sécurité, et bien sûr du niveau de maîtrise des technologies notamment de virtualisation.

L'objectif également de ces patterns est à la fois de simplifier le sujet, de limiter la variation des choix en matière de conception, et ainsi de favoriser les mutualisations des composants et des infrastructures. Ces patterns mettent avant tout en évidence les différents modèles de collaboration dans la mise à jour et l'accès aux données. Il n'y a pas de variation en tant que tel sur les composants intrinsèques au référentiel d'un pattern à l'autre. Le présent cadre n'a pas vocation à imposer une solution ou une technologie à ce stade, mais bien une compréhension, une vision, des règles, types d'architectures communes.

La logique générale de mise en place d'un référentiel et notamment de ses services répond aux principes « *API first* » et « *open first* » : à savoir, concevoir l'ensemble du dispositif sous une forme ouverte et accessible largement et simplement à travers des API utilisant les technologies et les principes d'interopérabilité du web (cf. RGI).

5.1. Architecture type d'un référentiel

L'architecture logique type d'un référentiel se présente selon la figure ci-après.

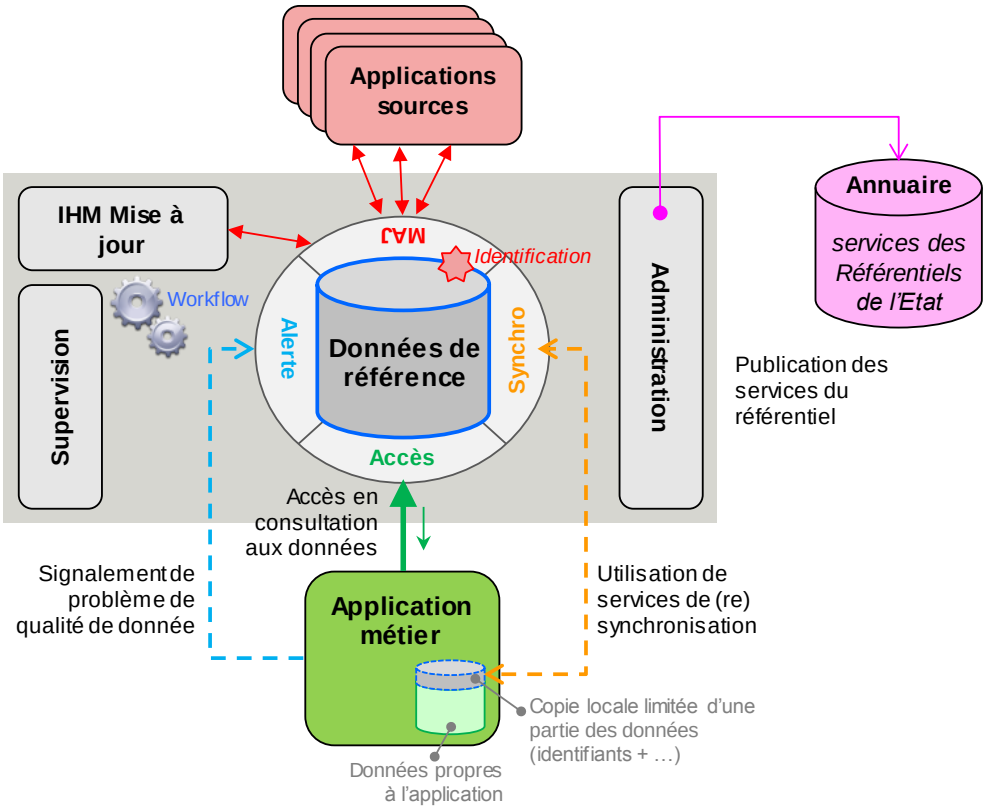

Figure 14 - Architecture logique type d'un référentiel

Elle comprend en son cœur un dispositif de stockage des données, par exemple une base de données (le plus souvent relationnelle, mais les technologies actuelles de type noSQL peuvent présenter des avantages significatifs dans certains cas de figure) accessible par une couche de service :

- Services de mise à jour (création, identification – c'est-à-dire d'attribution d'un identifiant, et mise à jour) ; Cet ensemble de services comprend également un service d'identification permettant d'attribuer à une nouvelle instance d'objet métier, un identifiant unique.
- Services d'accès (en lecture).

L'objectif sur ces deux premiers ensemble n'est pas de construire de simples services de type CRUD[23], mais bien un ensemble de services à valeur ajoutée pour les producteurs et les utilisateurs (consommateurs) des données de référence.

- Services d'alertes permettant de signaler des erreurs détectées en aval ; il est possible également d'intégrer des mécanismes de publication (type push) d'alerte à l'initiative du référentiel.
- Services de synchronisation ou de resynchronisation volontaire des données.

Elle comprend également plusieurs autres composants :

- Un composant permettant de gérer des workflows de mise à jour complexe, de gestion d'alerte…
- Un composant de supervision notamment de la qualité des données du référentiel. Les outils à proprement parler de gestion de la qualité des données ne sont pas intégrés à ce niveau, mais ils peuvent l'être dans certain cas.
- Un composant permettant si nécessaire de réaliser directement des IHM de mise à jour des données.
- Un composant général d'administration : modélisation des données, de gestion des services, de gestion de la sécurité des accès, etc.

Dans la présentation des patterns qui suivent, sont également identifiés dans les schémas :

- Les applications métiers (en rouge dans les schémas) en amont entrant dans les processus de création ou de mise à jour des données ;
- Les applications métiers utilisatrices du référentiel (en vert dans les schémas) ;
- Un« annuaire » des services : le lieu de consolidation et mise à disposition du catalogue interministériel des services d'accès aux données de référence du SI de l'État. Le terme « annuaire » n'est pas à prendre ici au sens technique, mais bien comme une liste de type « page jaune ».

Les services strictement liés aux échanges (type bus d'échange) entre les applications et le référentiel ne sont pas identifiés sur ces schémas. Les mécanismes de transport, de routage, de cache ne sont donc pas présentés dans ces patterns. Le but de ces schémas est bien d'identifier les différents modes de distribution et de collaboration des données de référence.

5.2. Pattern 1 : référentiel centralisé

Le premier pattern correspond au mode de distribution et de collaboration le plus simple. Les données de référence sont centralisées dans un référentiel unique, et leur mise à jour est effectuée directement dans le référentiel sans intermédiaire.

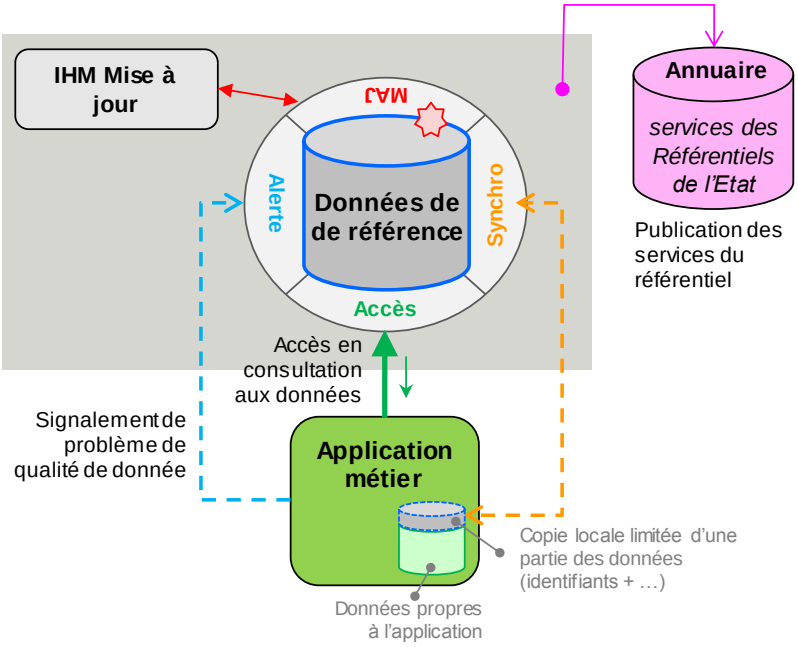

Figure 15 - Pattern 1 : Référentiel centralisé

Pour ce type de pattern, le niveau de maîtrise des processus de mise à jour des données est maximum, et donc le niveau de qualité de données est potentiellement le meilleur. Toutefois, ce modèle d'organisation et de coopération est difficilement applicable dans la pratique pour des données de référence complexes, comme par exemple les « Entreprises ». Mais il est clairement une cible à atteindre dans la plupart des cas.

Il est parfaitement adapté pour des nomenclatures, et en particulier des nomenclatures propres à l'administration, qui ne proviennent pas d'une source externe (par exemple internationale).

[23] CRUD : *Create, Read, Update, Delete*

La copie locale de tout ou partie des données de référence dans une application métier utilisatrice est a priori à exclure (cf. les règles RA3 et RA4). Dans le cas où elle est autorisée et donc encadrée par le RZF, l'application métier a la charge de s'assurer de la cohérence dans le temps de cette copie avec l'originale, en utilisant les services de resynchronisation volontaire mis à disposition par le référentiel.

5.3. Pattern 2 : Référentiel de consolidation

Le deuxième pattern introduit la possibilité de déporter la collecte, et la mise à jour des données de référence dans des applications métiers, la partie aval du dispositif ne change pas. Le couplage entre les applications sources et le référentiel est lâche, chacun reste autonome. Les processus de mises à jour ne sont pas vus globalement.

Dans ce pattern, il n'est pas précisé si les données de référence pouvaient ou non être stockées dans les applications sources (en générale, elles le seront). Cela n'a que peu d'importance, dans la mesure où le point de vérité identifié pour la donnée pour ce pattern est le référentiel. La question toutefois de l'éventuelle desynchronisation des données en cas de localisation amont reste ouvert, et les mécanismes de resynchronisation volontaire pourront également servir pour traiter ce point dans la chaîne amont du référentiel.

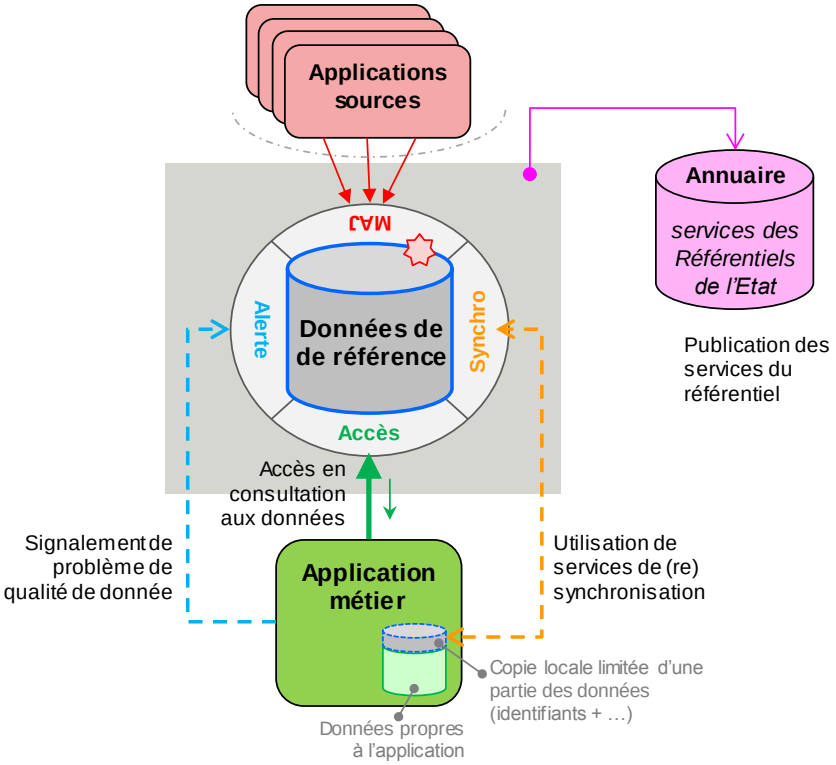

Figure 16 - Pattern 2 : Référentiel de consolidation

Ce pattern est généralement utile dans le cas de données de référence complexes, dans le sens, où de nombreux acteurs et processus contribuent à leurs mises à jour. C'est typiquement l'exemple des données de référence sur les Entreprises. Les objets métiers sous-jacents sont relativement riches et nombreux sémantiquement (unité légale, établissement, localisation, patrimoine, représentant, etc.). L'entretien de ces objets métiers met en œuvre un nombre d'acteurs significatifs, en particulier les Centre de Formalité des Entreprises (CFE) qui prennent différentes formes : chambre d'agriculture, chambre des métiers, greffes, centre des impôts, URSSAF… chacun de ces acteurs disposent de ses propres applications et gèrent ainsi une partie de la collecte et la mise à jour des données de référence qui sont ensuite consolidées par l'INSEE dans le référentiel SIRENE.

Le principal avantage de ce pattern est qu'il est relativement peu intrusif dans le SI, puisqu'il est possible de n'impacter que faiblement les applications sources. L'inconvénient majeur de ce pattern réside principalement dans sa capacité limitée à supporter les refontes ou évolutions de processus de collecte et de mises à jour des données, ou l'adaptation continue du modèle sémantique des objets métiers. Les applications source ne peuvent pas être considérées comme partie intégrante du dispositif, dans la mesure où le couplage est lâche entre les applications sources et le référentiel, et où les processus de mises à jour ne sont pas vus globalement.

Ce pattern peut présenter éventuellement des variantes en fonction de la localisation du service d'identification (qui attribue un identifiant). Il est possible d'envisager une attribution provisoire localement à une application source. Mais il est important de considérer que ce doit être une exception pour ce type de pattern (cf. le pattern 3).

La distribution amont de la collecte des données dans différentes applications doit clairement être définie : tout ou partie du périmètre fonctionnel, et/ou tout ou partie de la couverture du référentiel. Il ne doit bien évidemment pas y avoir d'intersection de périmètre ou de couverture dans les applications sources.

Dans ce pattern, on cherchera à simplifier au maximum le mode d'alimentation du référentiel, en évitant les variations par application source.

La copie locale de tout ou partie des données de référence dans une application métier utilisatrice est a priori à exclure (cf. les règles RA3 et RA4). Il est important dans ce pattern de bien considérer que les données contenues dans le référentiel sont en partie déjà des copies de données locales aux applications sources, avec les questions de désynchronisation sous-jacentes.

Dans le cas où elle est autorisée et donc encadrée par le RZF, l'application métier a la charge de s'assurer de la cohérence dans le temps de cette copie avec l'originale, en utilisant les services de resynchronisation volontaire mis à disposition par le référentiel.

5.4. Pattern 3 : Référentiel de coopération

Ce troisième pattern est une évolution du précédent. Les applications sources peuvent être considérées comme partie prenante de l'ensemble du dispositif : les processus sont globaux, et les applications sources utilisent elles-mêmes les données de référence. En réalité c'est bien l'ensemble (référentiel et applications sources) qui doit être considéré comme étant le « référentiel de données ». Cela signifie donc que dans ce cas, le responsable du référentiel a la capacité de faire évoluer les applications sources. Les processus d'entretien des données sont bien vue globalement.

Le couplage entre le référentiel et les applications sources est donc plus fort, et nécessitera généralement des outils d'intermédiation spécifiques (qui ne sont pas représentés ici) pour gérer la complexité des échanges et notamment de la synchronisation des données.

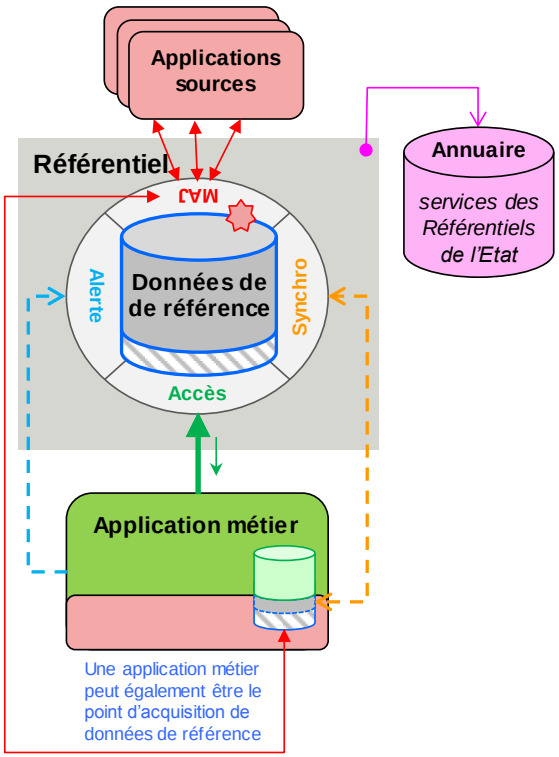

Figure 17 - Pattern 3 : Référentiel de coopération

L'avantage de ce pattern par rapport au pattern 2 est bien son agilité globalement supérieure. Il préserve toutefois moins l'existant, car il nécessite : des adaptations plus fortes du côté des applications sources, et un dispositif d'intermédiation pour gérer la synchronisation des données. C'est globalement un pattern plus complexe à mettre en place que le pattern précédent.

Comme dans le précédent pattern, la distribution amont de la collecte des données dans différentes applications doit clairement être définie : tout ou partie du périmètre fonctionnel, et/ou tout ou partie de la couverture du référentiel. Il ne doit bien évidemment pas y avoir d'intersection de périmètre ou de couverture dans les applications sources.

Dans ce pattern, il est possible d'imaginer une variation plus grande par application source du périmètre et de la couverture alimentée.

La copie locale de tout ou partie des données de référence dans une application métier utilisatrice est a priori à exclure (cf. les règles RA3 et RA4). Dans le cas où elle est autorisée et donc encadrée par le RZF, l'application métier a la charge de s'assurer de la cohérence dans le temps de cette copie avec l'originale, en utilisant les services de resynchronisation volontaire mis à disposition par le référentiel.

5.5. Pattern 4 : Référentiel esclave

Ce pattern introduit un autre aspect dans l'architecture de collecte, de stockage et de distribution des données de référence. Il traite le cas de données de référence qui, pour des raisons historiques, ou pour des raisons intrinsèques à leur nature, s'appuient sur d'autres données de référence gérées dans un autre référentiel, appelé « référentiel maître » ou « référentiel socle ». C'est par exemple le cas de données de référence sur les carrières qui s'appuient sur des données de référence à la fois sur les entreprises et les personnes, ou encore, les données fiscales sur les entreprises qui s'appuient bien évidemment sur les données d'identification des entreprises. Ce principe est la déclinaison applicative de la règle RF5 concernant les dépendances entre données de référence.

Ce pattern peut se décliner selon différentes variantes (cas des patterns 1 à 3) sur l'amont du référentiel (collecte et mise à jour des données) qui ne sont pas représentées ici.

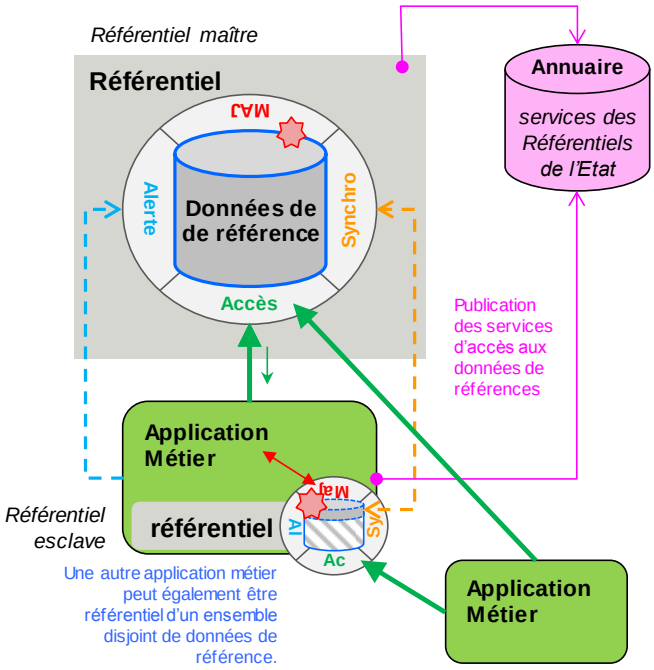

Figure 18 - Pattern 4 : Référentiel esclave

La principale difficulté, ou contrainte, dans la mise en place de ce pattern consiste à limiter au maximum la duplication des données du référentiel maître vers le référentiel esclave pour éviter tout problème d'incohérence. Dans tous les cas, les applications utilisatrices du référentiel esclave qui souhaitent accéder aux données de référence du référentiel maître devront systématiquement appeler directement le référentiel maître (pour limiter la réutilisation de données déjà recopiées).

5.6. Pattern 5 : Référentiel hub

Le dernier pattern identifié, appelé référentiel hub, consiste à virtualiser la consolidation des données de référence. Il se rapproche le plus du pattern 2. Le point de vérité reste en réalité dans chaque application source.

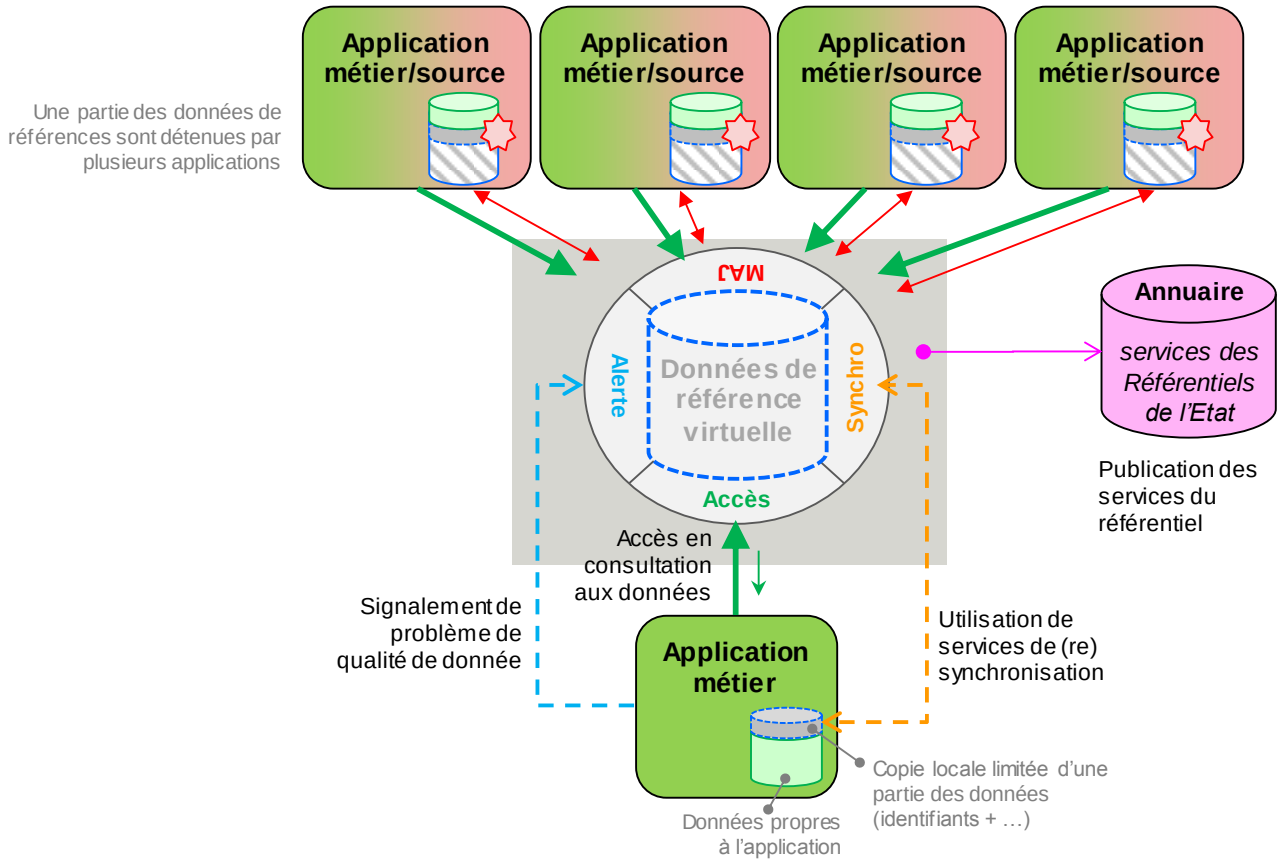

Figure 19 - Pattern 5 : Référentiel hub

Ce type de pattern nécessite une architecture d'échange particulière de type EII/EAI/ESB. Comme le point de vérité reste propre à chaque application source, et donc propre à chaque sous ensemble de données de référence, ce type de pattern n'est préconisé que pour des données de référence qui peuvent être structurées et organisées aisément (ce qui est très rarement le cas), et que pour des usages de type consommation massive.

6. MISE EN PLACE & GOUVERNANCE

6.1. Les étapes de mise en place d'un référentiel

Le présent cadre propose une démarche de mise en place d'un référentiel. La figure ci-après illustre cette démarche. Cette démarche ne doit pas être considérée comme un cadre rigide mais uniquement comme un guide synthétique de mise en place. Ce chapitre fera l'objet de complément ou de précision en fonction des premières utilisations, selon le principe d'amélioration continue.

La numérotation des étapes décrit la logique d'ensemble et de dépendance des travaux à conduire. Cette démarche peut d'ailleurs être utilisée dans les évolutions successives d'un référentiel, sous la forme d'un cycle, chaque itération étant une version du référentiel ou une version de la conception du référentiel.

Le niveau de précision de chaque étape dépend donc avant tout de la trajectoire de mise en place du référentiel. La première version pouvant être limitée, et enrichie successivement sur tel ou tel aspect (couverture des données, complétude des services, performance, etc.). L'ordre des étapes est indicatif. Les couleurs signalent les étapes qui sont liées entre-elles et/ou qui sont de nature équivalente.

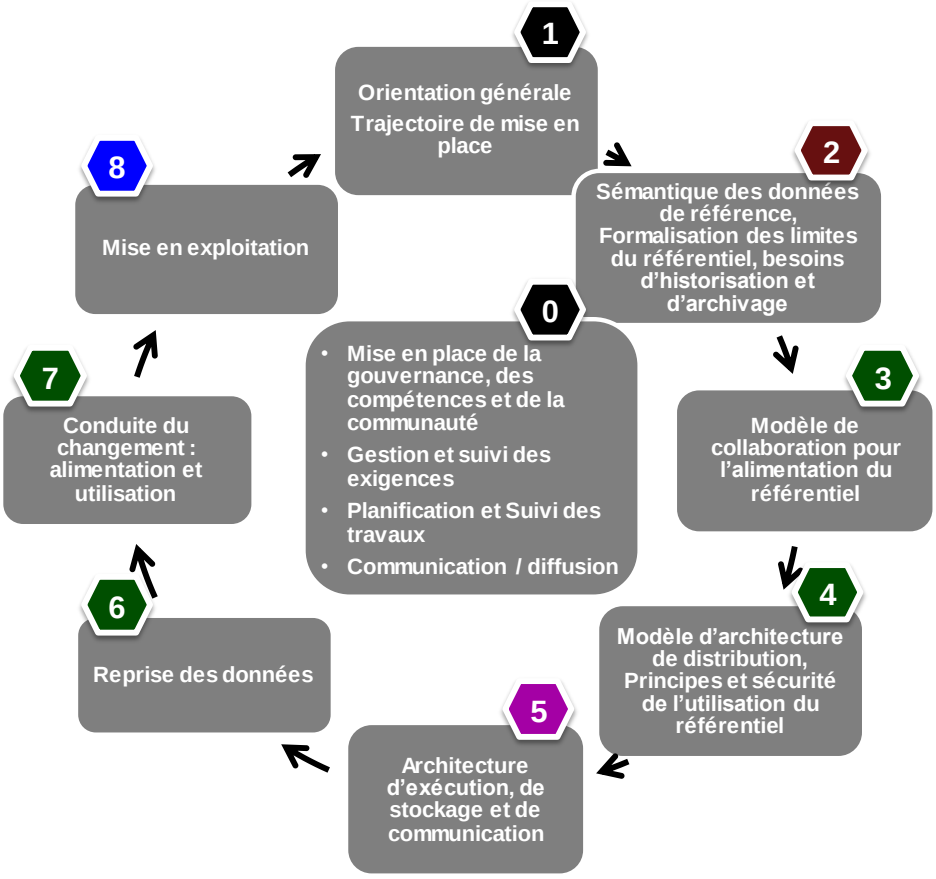

Figure 20 - Le cycle général et les étapes de mise en place d'un référentiel

Etape 0 : elle consiste à définir l'organisation, les compétences, le mode de pilotage nécessaires à la mise en place du référentiel. C'est à ce niveau que doit être défini et mis en place la comitologie, associant les services métiers et services informatiques, mais aussi les producteurs de données et consommateurs, donc utilisateurs du référentiel. Les services en charge des archives doivent également être identifié et impliqué en amont.

La gestion dans le temps des exigences est également un élément clé qui doit être identifié très en amont, même si sa mise en place peut être itérative. Enfin la nécessaire planification des travaux, et communication sur leur avancement sont des activités clés dans la mise en place d'un référentiel, qui est par nature un élément de partage et de communication au sein du SI.

Etape 1 : Il est nécessaire de consolider au plus tôt l'ensemble des éléments et des orientations générales permettant de cadrer le périmètre fonctionnel du référentiel, l'architecture générale (métier, fonctionnelle, et technique), les modalités de mise en place, les moyens alloués, et l'organisation nécessaire à son maintien en condition opérationnelle. Ce travail doit permettre de définir une trajectoire de mise en place du référentiel. Un référentiel de données étant un élément clé du SI, cette trajectoire doit s'intégrer de manière explicite et cohérente dans la trajectoire interministérielle de transformation du SI de l'État. C'est à cette étape, que doit être réalisé une première analyse sommaire de l'existant : quelles sont les données de référence existantes, les référentiels existants, leurs processus d'alimentation.

Etape 2 : La sémantique des données de référence doit être décrite très précisément et formalisée dans des modèles UML. Cette sémantique comprend la vision statique, dynamique et les règles de gestion associées. Cette étape doit être considérée comme incontournable, un effort tout particulier doit donc lui être réservé.

Cette étape doit également permettre d'identifier les besoins et les règles concernant le cycle de vie des données de référence, et d'amorcer l'analyse et la conception de la solution sur ces questions : besoin d'historisation, durée de conservation, stratégie d'archivage. Ce travail doit permettre d'identifier un premier ensemble de métadonnées associées aux données de référence (dictionnaire sémantique). Il sera complété sur les étapes 3 et 4.

Etape 3 : Le dispositif organisationnel, métier et fonctionnel permettant d'alimenter le référentiel de données : création et mise à jour des données, doit être conçu notamment à partir de la vision dynamique de la sémantique, à savoir le cycle de vie des objets métiers. Dans le cas de référentiel transverses à l'ensemble de l'État, ces dispositifs, ou modèles de collaboration pour l'alimentation, doivent être pensés et conçus de manière industrielle et automatisée dès la première mise en place. Il s'agit de concevoir ce modèle d'alimentation, mais également de préparer la trajectoire de mise en place de l'organisation métier nécessaire et la conduite du changement associée (étape 7). La qualité des données étant un élément critique dans l'utilisation de données de référence, c'est à cette étape que doivent être définis les indicateurs de qualité et leur mode de calcul et d'alimentation.

Ce modèle de collaboration doit s'appuyer sur l'un des patterns génériques proposés dans le présent guide, où éventuellement une combinaison de ces patterns.

Etape 4 : L'architecture de distribution des données dans le SI et donc les principes d'utilisation du référentiel doivent s'appuyer sur l'un des patterns du présent guide, ou une combinaison de ces patterns, avec deux orientations majeurs : « API » et « Open ». La conception des services d'accès aux données et notamment la mise à disposition de ces services de manière ouverte est au cœur de cette étape, avec comme exigences : la performance, l'engagement et la qualité de service, et la sécurité.

La complexité de cette étape réside dans la mise à disposition de services permettant de synchroniser ou resynchroniser les données de référence copiées dans des applications utilisatrices.

Cette étape, ainsi que la précédente, doit également permettre de compléter l'identification et la définition des métadonnées à mettre en place : en fonction du processus d'alimentation et du mode de distribution et notamment des besoins de traçabilité.

Etape 5 : La mise à disposition de services (par exemple des services web de type REST ou SOAP) sur un référentiel de données au sein d'un ministère ou au niveau même de l'État, nécessite absolument une infrastructure qui puisse monter en performance (scalable). Comme il l'a déjà été dit, il n'est pas raisonnable d'imaginer maîtriser tous les besoins et utilisations possibles de ces référentiels. Il est donc nécessaire de superviser très précisément l'utilisation du référentiel, et de mettre en place une architecture qui permette la scalabilité.

Note : L'évolutivité des infrastructures techniques de communication, d'exécution et de stockage, est moins une problématique à l'heure actuelle avec les technologies de virtualisation, qui « assurent » une certaine linéarité dans la montée en charge.

Etape 6 : Cette étape consiste à concevoir, planifier et réaliser la reprise des données, ou l'initialisation du référentiel, c'est-à-dire son premier chargement. Cette étape peut se révéler d'une très grande complexité, et se dérouler sur des délais importants. Elle est critique et ne doit pas être sous évaluée. La constitution d'une première « vraie » version peut nécessiter plusieurs mois voire plusieurs années, car elle nécessite notamment une réelle transformation des sources et des processus de collecte. Cette étape d'initialisation ou de chargement peut prendre différentes formes :

- Une initialisation en une seule étape
- Une initialisation au fil de l'eau : l'initialisation est réalisée progressivement par les processus de collecte mis en place. Il peut être nécessaire de faire un pré-chargement à la première mise en place
- Une initialisation par palier : premier chargement, suivi d'une mise en service (sur un périmètre, une couverture et une portée limitée), ce qui laisse le temps de préparer et de conduire le changement du deuxième chargement, sur un périmètre / couverture / portée plus large.

Cette étape comprend :

- L'étude des données dans les systèmes identifiés comme source, ou comme préfigurateur du référentiel.
- L'établissement de la correspondance entre les modèles des systèmes existants et le modèle du référentiel cible défini dans l'étape 2.
- L'alignement de la qualité des données existantes sur les attentes a minima du référentiel cible : transformation, transcodage, réalignement sémantique, réconciliation… Cette étape peut donc nécessiter la réalisation d'outils spécifiques de transformation de données et de chargement de données.
- Le chargement du référentiel.
- L'analyse et le traitement manuel ou automatique des rejets lors du chargement (avec autant d'itération que nécessaire).
- Et enfin la recette métier.

Etape 7 : Cette étape consiste à connecter le référentiel au reste du SI, et plus globalement à conduire le changement.

Même si l'intégration du référentiel dans le SI est une action plutôt de long terme : migration de toutes les applications métiers sur les services du référentiel. Il est impératif d'embarquer dès la première mise en service un nombre significatif d'applications métiers. La coordination de ces mises en place introduit souvent une complexité de planification et de coordination dans la mise en place des premiers services sur le référentiel. Il ne faut pas sous-estimer ce point.

La conduite du changement est une étape fondamentale. Il existe des méthodes et des pratiques éprouvées sur ce sujet. L'objet ici n'est pas de rentrer dans le détail, mais bien d'attirer l'attention sur la nécessité d'identifier et de traiter ce point. L'objectif est bien d'identifier quels sont les changements sur les organisations, les compétences, les modes de fonctionnement, les processus et procédures, les responsabilités, les outils, les personnes elles-même qui sont à organiser, à planifier et à suivre. La communication et la formation est notamment un point essentiel de ce travail.

Etape 8 : Cette étape consiste à préparer et mettre en place la gouvernance, la gestion et le maintien en condition opérationnelle (MCO) des données de référence et du référentiel en tant que tel (qui est décrit dans la paragraphe 6.2). Il est important dans la mise en place d'un référentiel d'anticiper le « run » du référentiel, c'est-à-dire d'identifier tous les dispositifs nécessaires organisationnels, techniques et les outils supportant les activités liées au fonctionnement en mode récurrent du référentiel.

6.2. Les activités de gouvernance, de gestion et de MCO d'un référentiel

Les activités de pilotage, de maintenance et de gestion dans le temps d'un référentiel sont illustrées dans la figure suivante. Il s'agit d'un premier guide et non d'un cadre rigide. Ce paragraphe aura l'occasion d'être complété et enrichi en fonction des premiers retours. Il reste volontairement succinct pour cette première version. Il s'agit à ce stade d'identifier les macro-activités concernées.

Figure 21 - Les activités de gouvernance et de gestion d'un référentiel

Cinq macro-activités doivent être mise en place et articulés autour d'une macro-activité de communication, de gouvernance et de supervision (suivi général des activités) :

- Une macro-activité de Gestion de la relation avec les utilisateurs : il s'agit classiquement de gérer la communication vers les utilisateurs du référentiel, de suivre les engagements de services, et de traiter et suivre toutes les demandes de ces derniers. Il est noté ici que la notion d'utilisateurs doit évidemment être prise au sens le plus large possible, qu'ils soient internes ou externes à l'État ou même la sphère publique.
- Une macro-activité de MCO, Maintien en Condition Opérationnelle : qui consiste à exploiter le référentiel de données à proprement parler et tous les dispositifs techniques et métiers associés qui touchent tant à l'alimentation du référentiel qu'à son utilisation. Cette macro-activité comprend également toutes les activités de maintenance (corrective et évolutive) et d'évolution du référentiel.
- Une macro-activité de Gestion de la qualité des données : qui consiste d'une part à mesurer et suivre la qualité des données, notamment grâce à des indicateurs publiés, ainsi que d'autre part au traitement de toutes les anomalies détectées par la supervision, par l'analyse des indicateurs, ou par les utilisateurs eux-mêmes.
- Une macro-activité d'Administration de la sémantique des données : il s'agit de gérer dans le temps la définition sémantique des données gérées dans le référentiel, leur traduction en termes de structures physiques de données (interne au référentiel, et surtout, dans la structure des données échangées dans les services).
- Une macro-activité d'Etude et d'architecture : elle comprend toutes les études amont de définition, de mise en place, d'évolution ou transformation du référentiel : son architecture métier, fonctionnelle, applicative ainsi que l'infrastructure de communication, d'exécution, de stockage ou de service nécessaire.
- Une macro-activité enfin de Communication, de Gouvernance et de Supervision : elle a le rôle de pilotage d'ensemble tant opérationnel que stratégique sur le long terme. Elle doit veiller, et c'est un point capital pour un référentiel de données, à une communication la plus précise, actualisée, large et ouverte possible. Des référentiels comme celui des Entreprises ou des Adresses sont autant utilisés au sein de l'État que par le secteur privé : la communication sur la qualité des données, les services disponibles et leurs SLA, mais aussi la trajectoire de transformation du référentiel doit être disponible directement sur internet de manière ouverte (principe d'open API).

Les versions suivantes de ce cadre préciseront ce premier cadre d'activité.

A CHECK-LIST DE LA CONFORMITÉ AU CADRE

Dans le cadre de l'évaluation de la conformité de référentiels de données existants au présent cadre, il a été jugé utile de définir une check-list sommaire identifiant l'ensemble des points à vérifier.

L'objectif de cette check-list est :

- d'évaluer selon une même grille d'analyse les référentiels de données majeurs du SI de l'État.
- d'analyser les forces et faiblesses de ces référentiels et des dispositifs techniques et métiers qui les accompagnent
- de définir et piloter la trajectoire d'alignement des référentiels sur ce cadre, et sur les évolutions métiers attendues.

Cette check-list est illustrée par la figure suivante.

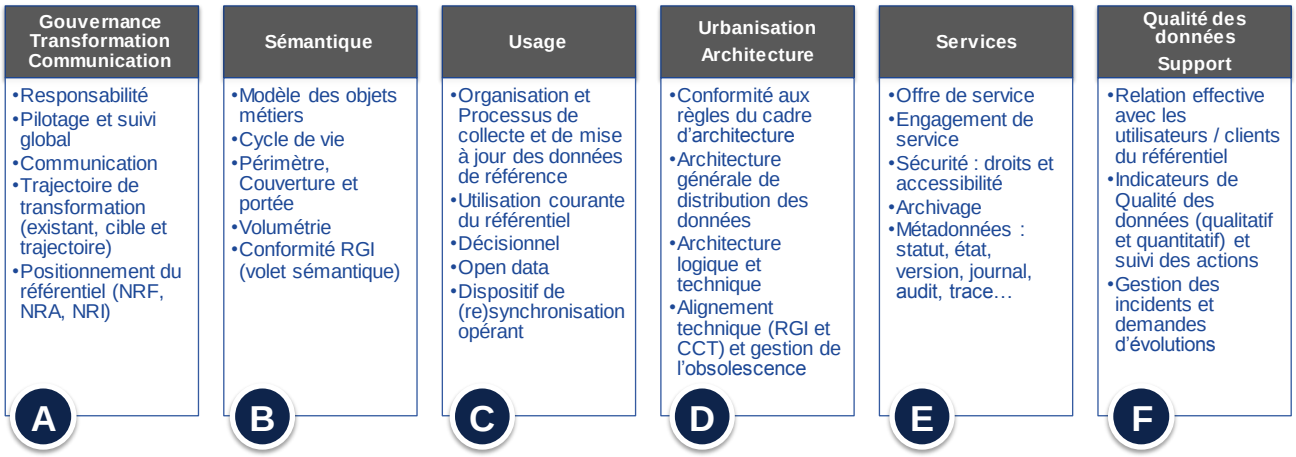

Figure 22 - Check-list de conformité au présent cadre

Elle s'utilise avec la grille de maturité provenant du modèle CMMi[24] qui comprend les 5 niveaux suivants :

- Niveau 1, ou Initial : le résultat final est imprévisible ;
- Niveau 2, ou Discipliné : l'activité ou le dispositif est structuré et reproductible ;
- Niveau 3, ou Défini ou Ajusté : l'activité ou le dispositif est standardisé et institutionnalisé ;
- Niveau 4, ou Contrôlé ou Géré quantitativement : l'activité ou le dispositif est mesurée et planifiée, la qualité et la criticité sont prises en compte, un premier niveau de pilotage de la performance est en place ;
- Niveau 5, ou Optimisé : l'activité ou le dispositif est totalement décrit, maîtrisé et en optimisation constante.

Une première utilisation consiste à évaluer globalement le dispositif dans son ensemble avec une seule note de maturité.

Une deuxième utilisation consiste à évaluer de manière macroscopique le dispositif sur chacun des six thèmes identifiés (de A à F), chacun évalué avec une note globale de maturité.

Et enfin, une troisième et dernière utilisation de cette check-list consiste à évaluer chaque point de chaque thème. Une combinaison de la deuxième et de la troisième utilisation est bien sûr envisageable.

NdA : les termes NRF, NRA, NRI désignent les nomenclatures de référence fonctionnelle, applicative et infrastructure. Se référer au Cadre Commun d'Urbanisation du SI de l'Etat[25].

[24] CMMi : Capability Maturity Model Integration,
[25] http://references.modernisation.gouv.fr/urbanisation-du-systeme-dinformation-de-letat

B MODÈLE DE CATALOGUE DE RÉFÉRENTIEL ET DE SERVICES ASSOCIÉS

En dehors du Cadre commun d'urbanisation du SI de l'État qui précise le métamodèle applicable pour la gestion de la connaissance, et notamment la connaissance sur les actifs de type « Application », « Objet métier » et « Service », la présente annexe précise les attributs de ces objets.

Un référentiel de données est une application, certes particulière, mais cela reste une application à part entière.

	Attribut	Définition	Format	Multiplicité
Application	Secteur	Domaine/Zone/Quartier/Bloc du POS du SI de l'Etat couverts par l'application (il peut y en avoir plusieurs)	Nom	1..*
	Autorité Administrative	Nom du ministère (sigle) ou de l'Administration responsable	Sigle	1
	Nom	Nom de l'application (nom partagé entre les différents acteurs)	Nom	1
	Code	Code d'identification de l'application utilisé par la production	Code alphanumérique	0..1
	Etat	Etat de l'application, par rapport à son cycle de vie : I = identifiée dans la trajectoire C = en construction, P = en production, R = en retrait	I, C, P, R	1
	Type	Type d'application : M = Application Métier, B = Outil bureautique, R = Référentiel de données, E = Outil d'exploitation, D = outil de développement	M, B, R, E, D	1
	Nature	Nature de l'application : L = Logiciel libre, P = Progiciel, S = Spécifique, M = Combinaison de développement spécifique et/ou de progiciel et/ou de logiciel libre	L, P, S, M	1
	début d'exploitation	Année de début d'exploitation de l'application (prévisionnelle pour les applications en construction, réelle pour les application en production)	Année	1
	fin d'exploitation	Année de fin d'exploitation de l'application (prévisionnelle pour les applications en production)	Année	0..1
	Description	Description synthétique des services rendus par l'application	Texte	1
	Cible	Qualification par la DSI de la pérennité de l'application (obsolescence technique, fonctionnelle…) : O = Application Cible ou pérenne et positionnée dans la trajectoire, N = application dont le retrait est planifié ou doit l'être pour des raisons techniques et/ou métiers	O, N	1
	Sensible	identifie si l'application est considérée comme sensible ou non : utiliser dans des processus critiques, contenant des données sensibles…	O, N	1
	Direction métier	Nom de la direction métier principale	Nom	1
	Responsable métier	Nom, Prénom, Email du responsable métier (MOA)	Nom+Prénom+Email	0..1
	Responsable SI	Nom, Prénom, Email du responsable technique (MOE)	Nom+Prénom+Email	0..1
	Nombre d'utilisateur	Estimation du nombre d'utilisateurs	Entier	0..1
	Objets métiers	Liste des principales données de références et/ou nomenclatures disponibles	Liste d'objets métiers	0..*
	Structure de données	Identifie si la structure des données est disponible et accessible, présence de modèle de données (MPD)	O, N	1
	Services d'accès	Identifie si les données et nomenclatures sont accessibles par des services applicatifs publics	O, N	1
	Organisation des données	Identifie le mode de d'organisation physique des données (stockage) indépendemment de leur gestion métier : C = Centralisé, L = Localisé, M = Mixte	C, L, M	0..1
	Qualité	Description succincte du niveau de qualité des données (complétude, cohérence, fraicheur, précision, doublon,…)	Texte	0..1
	Niveau de maturité	Identification du niveau de maturité de gestion de l'information : de 1 à 5 (cf. EIM maturity model du Gartner)	1, 2, 3, 4, 5	0..1
	Disponibilité	Niveau de disponibilité (SLA) de l'application	Texte	0..1
	CNIL	Présence de données personnelles soumises à la CNIL	O, N	1

Figure 23 - Structure de données d'une Application de type référentiel

	Attribut	Définition	Format	Multiplicité
Objet métier	**Secteur**	Domaine/Zone/Quartier/Bloc du POS du SI de l'Etat dans lequel est positionné l'objet (il ne doit y en avoir qu'un seul)	Nom	1
	Autorité Administrative	Nom du ministère (sigle) et de l'Autorité Administrative responsable / référente	Sigle	1
	Nom	Nom de l'objet métier (le nom de l'objet métier doit être unique)	Nom	1
	Description	Description synthétique de l'objet métier	Texte	1
	Structure	Identifie si la structure des données est disponible et accessible, présence de modèle de données (équivalent diagramme de classe UML) : O/N	O, N	0..1
	Cycle de vie	Identifie si le cycle de vie de l'objet métier est défini, disponible et accessible (équivalent diagramme d'état UML) : O / N	O,N	0..1
	Référence	Précise le mode d'organisation de la gestion de la donnée (porte bien sur les processus de gestion : création, modification, suppression, indépendemment du stockage, qui lui est porté par l'application) : C = totalement Centralisé, L = Localement dans différentes structures LC = Localement avec une consolidation centralisée M = Mixte, pour tous les autre cas, comme par exemple une gestion centralisée, mais avec des compléments apportés à un niveau local	C, L, LC, M	0..1
	Qualité	Description succincte du niveau de qualité des données (unicité, complétude, exactitude, actualité, conformité, intégrité, cohérence, accessibilité, pertinence) et notamment si les processus de création, modification, et suppression sont identifiés, formalisés et sous contrôles (outillés, indicateurs…).	Texte	0..1
	Application	Nom de l'application dans laquelle le point de vérité pour cet objet est disponible	Nom	1
	Décisionnel	Cet objet métier est-il utilisé également dans les applications de type « décisionnel » (entrepôt, DWH, Datamart, etc.) : O/N	O,N	0..1

Figure 24 - Structure de données d'un Objet métier

	Attribut	Définition	Format	Multiplicité
Attribut (Donnée)	**Objet métier**	Nom de l'objet métier caractérisé par l'attribut ou la donnée	Nom	1
	Nom	Nom en clair de l'attribut (le nom de l'attribut doit être unique pour	Nom	1
	Code	code ou libellé technique de l'attribut	Texte	0..1
	Multiplicité	Indique le nombre d'occurrence possible (par défaut 1)	1,0..1,0..*,1..*	0..1
	Type	Nature ou type de l'attribut. Le typage peut-être technique ou métier. Dans le cas d'un typage métier (à partir d'un dictionnaire de type), la référence du dictionnaire de type doit être indiquée.	Nom	1
	Taille	Si nécessaire (en complément du type) la taille de l'attribut peut être indiquée	Entier	0..1
	Description	Texte décrivant le sens de l'attribut, dans un langage et un vocabulaire compréhensible et si possible autoporteur	Texte	1
	URL définition	Lien url vers laquelle il possible de retrouver la définition actualisée lorsqu'elle existe	URL	0..1
	Valeurs	Liste des valeurs possibles de l'attribut	Liste de Nom	0..*
	Standard	Indique la référence du standard correspondant à l'attribut (ex. UN/CEFACT/CC/Person.Name)	URL ou Texte	0..*
	Dictionnaire	Indique la référence (si possible une url) du dictionnaire de type de données métiers utilisé	URL ou Texte	0..1

Figure 25 - Structure de données d'un Attribut d'un objet métier

	Attribut	Définition	Format	Multiplicité
Services	**Application détentrice**	Nom de l'application détentrice du service	Nom	1
	Nom	Nom ou code du service	Nom	1
	Description	Description succincte du service, notamment : - des modes d'accès aux données (synchrones / asynchrones) - du type d'action (lecture / modification) - unitaire / lot	Texte	1
	Objets métiers	Liste des principales données de références et/ou nomenclatures disponibles	Liste d'objets métiers	0..*
	Type	Définie le type de service : service web, REST…	Texte	1
	Service public	Définie si le service est accessible depuis une API ouverte	O, N	1
	Version	Dernière version disponible	Texte	0..1
	URL	Adresse URL du service	URL	0..1
	Documentation	URL de la documentation technique du service (interface d'accès, fichiers xml)	URL	0..1

Figure 26 - Structure de données d'un Service

C GLOSSAIRE DE LA GOUVERNANCE DES DONNÉES

La présente annexe recense l'ensemble des termes utilisés et la définition retenue au niveau interministériel pour la gouvernance des données de l'État. Dans plusieurs cas, le contexte dans lequel s'applique cette définition est précisé entre crochet.

Accessibilité [Critère de qualité de la donnée]
Actualité [Critère de qualité de la donnée]
Bloc (fonctionnel) [Urbanisation du SI de l'État]
Clé [Architecture SI]
Cohérence [Critère de qualité de la donnée]
Complétude [Critère de qualité de la donnée]
Confidentialité [Critère de qualité de la donnée]
Conformité [Critère de qualité de la donnée]
Critère de qualité
Cycle de vie (d'une donnée)
Donnée
Donnée brute
Donnée agrégée
Donnée structurée
Donnée semi-structurée
Donnée non-structurée
Donnée de référence
Donnée de contexte
Donnée maître
Donnée constitutive
Donnée de paramètre
Documents
Disponibilité [Critère de qualité de la donnée]
État (diagramme d'état)
Exactitude [Critère de qualité de la donnée]
Folksonomie
Fonctionnalité
Historisation
Identification
Identifiant
Intégrité [Critère de qualité de la donnée]
Métadonnée
Modèle de données
Nomenclature
Nomenclature de référence [Urbanisation du SI de l'État]
Objet métier
Ontologie
Pertinence [Critère de qualité de la donnée]
Plan d'Occupation des Sols (du SI de l'État)
Point d'acquisition
Point de vérité
Point de consommation

Quartier (fonctionnel) [Urbanisation du SI de l'État]
Référentiel / Référentiel de données
Référentiel de nomenclature
Référentiel local
Référentiel principal
Référentiel réglementaire
Répertoire
Sémantique
Service d'accès
Structure de données
Taxonomie
Thesaurus
Traçabilité [Critère de qualité de la donnée]
Type de données
Unicité [Critère de qualité de la donnée]
Versionning
XML (langage)
XML (schema)
Zone (fonctionnelle) [Urbanisation du SI de l'État]