Prescriptivité Référentielle[1]

Une approche systématique de l'excellence référentielle

1 Introduction

Ce mini livre blanc introduit les concepts et fonctionnalités génériques qui sont nécessaires pour qu'un référentiel de données soutienne efficacement un système d'information centré sur les données. Les aspects méthodologiques et organisationnels de la gouvernance des données étant bien documentés, ce papier se focalise sur les aspects fonctionnels et technologiques du sujet.

2 Un défi complexe

La numérisation et l'interopérabilité rendent les référentiels toujours plus pertinents: personnes, organisations, lieux, bâtiments, produits, catégories et nomenclatures, toutes ces informations doivent être partagées dans l'ensemble du système d'information.

Construire un système robuste et fiable se révèle, néanmoins, extrêmement difficile.

Les facteurs de complexité incluent :

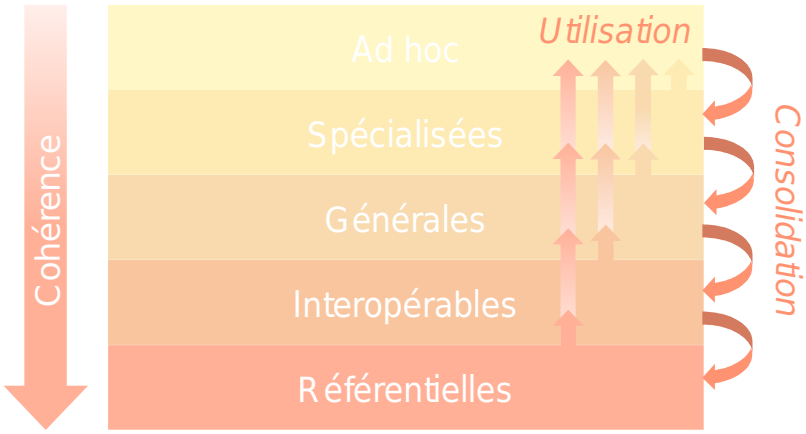
- Les volumes et la fréquence des mises à jour;
- Le nombre et la diversité des sources ;
- La diversité des données (concepts, intégrité, périmètre, granularité temporelle, niveau de détail...) ;
- La variabilité des niveaux de qualité et d'actualité des données ;
- La difficulté d'identifier les entités et de convertir les données vers des bases communes ;
- Les statuts et cycles de vie des données, leur intégrité et leurs interdépendances, les processus asynchrones ;
- Les valeurs conflictuelles, les exceptions, les valeurs contextuelles ;
- Les exigences de sécurité, y compris le contrôle fin des accès et la conformité à la réglementation sur la protection des données personnelles.

La complexité tend à s'accroître, du fait de l'évolution de ces facteurs: la complexité des systèmes augmente avec les volumes, la couverture fonctionnelle et les besoins d'intégration, cependant que la législation devient de plus en plus exigeante et les menaces contre la sécurité de plus en plus sophistiquées.

3 Modèle de cohérence des données

Les données appartiennent à des catégories très différentes: certaines sont structurantes pour l'ensemble de l'organisation, d'autres s'appliquent à un périmètre très étroit ; certaines sont durables, d'autres ont une pertinence limitée à une courte période.

L'urbanisme du système d'information et la technologie du référentiel doivent prendre en compte cette diversité. En particulier, ils doivent permettre, et même soutenir efficacement la stratification naturelle des données décrite dans le schéma suivant :



- Au sommet de la pile des données se tiennent les **données ad hoc**, lesquelles peuvent intégrer des informations d'autres couches, mais ne sont pas techniquement des sources pour le reste du système d'information ; leur volatilité n'est de ce fait pas limitée. Ces données sont souvent dans un format orienté utilisateur, notamment bureautique, tel que feuilles de calcul ou présentations, courriels ou simplement messages instantanés, audio ou vidéo ;
- Juste en-dessous dans le système se trouvent les **données spécialisées** ; ces données restent locales à la partie du système où elles sont produites, i.e. limitées soit à une partie de l'organisation (e.g. données de contrôle industriel) ou à quelque fonction transversale dont l'intégration est limitée (e.g. entrées de calendrier vCal);
- Plus bas siègent les **données générales**: bien que n'étant pas intégrées à l'ensemble de l'organisation, ces données sont pertinentes au-delà de leur périmètre d'origine, et méritent d'être régulièrement[2] rapprochées, comparées ou agrégées dans quelque outil de contrôle ou d'analyse ; ces données doivent être suffisamment cohérentes pour être utilisables dans ces consolidations;
- Plus profondément dans le système, les **données interopérables** au sens strict deviennent nécessaires pour coordonner les tâches à travers l'ensemble du système : par exemple, les détails d'une commande doivent être transmis scrupuleusement et compris exactement à chaque étape pour être traités à travers la production, l'expédition et la facturation ;
- Au niveau fondamental se tiennent les **données de référence** ; ces données sont centrales pour plusieurs domaines, ce qui requiert leur solidité et stabilité ; quand la qualité des données de référence (pertinence, exactitude, exhaustivité) est suffisante, elles peuvent être utilisées de façon prescriptive, et structurer les processus dans l'organisation, et même au-delà (e.g. catalogue produits).

Les données sont utilisées facilement et sans surcoût au niveau auquel elles appartiennent et aux niveaux supérieurs ; consolider les données d'un niveau à un niveau inférieur, au contraire, exige un effort spécifique d'organisation des données, à travers quelque combinaison des moyens suivants :

---

2 'régulièrement' est important ici : il est bien sûr possible de tirer des statistiques à partir du contenu de courriels, de rendez-vous ou de données volatiles de capteurs industriels, par exemple ; néanmoins, de telles études ne sont pas ordinaires, et leurs résultats appartiennent typiquement à la couche de données ad-hoc.

- Emploi d'heuristiques : règles d'induction, analyse statistique (big data) ;
- Apport de valeur ajoutée : intendance des données (data stewardship), gestion de la qualité des données (DQM).

Cette asymétrie entre utilisation et consolidation des données est non seulement inévitable, mais encore solidement fondée en théorie de l'information : comme en thermodynamique, la qualité (exactitude, homogénéité, complétude…) des données ordonnées[3] à leur emploi (pertinence, disponibilité, actualité…) n'apparaît pas spontanément ; un effort (recoupements, analyse des corrélations, enquêtes, gouvernance[4]) doit donc être consenti pour l'augmenter.

4 Les principes de la prescriptivité

L'objet même des données de référence est d'être utilisées systématiquement lorsqu'elles s'appliquent, afin d'assurer la cohérence et l'efficacité des échanges de données et des processus correspondants.

L'excellence est données est un prérequis de ce rôle prescriptif: une prescriptivité décrétée sur des données douteuse échouera, dans le meilleur des cas ; appliquée aveuglément, elle pourrait même dégrader la pertinence des données de l'organisation et perturber ses processus.

Tandis que l'approche statistique big data est suffisante tant que la prescriptivité n'est pas en jeu, dès que les données non seulement guident mais encore structurent les processus métiers, l'exactitude devient une nécessité.

Cette excellence des données ne pose pas de défi particulier s'agissant de concepts clairement définis, ayant une matérialisation assez stable dans le temps, surtout lorsque qu'une source unique et cohérente est disponible ; ainsi en va-t-il, par exemple, des codes postaux et des fuseaux horaires.

À l'inverse, obtenir l'excellence des données qui va permettre la prescriptivité dans des domaines où la complexité est grande nécessite un investissement significatif, organisationnels d'une part, et techniques de l'autre :

- Sans les processus et la formation centrées sur les données, les meilleurs outils resteront impuissants, ne pouvant créer les informations manquantes[5];
- Réciproquement, sans des mécanismes efficaces d'intégration de données, sans la possibilité d'exprimer la richesse de l'information[6], la gestion de la qualité des données restera trop coûteuse pour atteindre une cohérence satisfaisante des données de l'organisation.

Plus spécifiquement, une modélisation et une gestion de la qualité rigoureuses sont nécessaires pour produire un bon niveau d'exactitude et d'exhaustivité des données, donc la pertinence du référentiel en tant que prescripteur:

- Le modèle de données du référentiel doit être suffisamment puissant pour refléter la complexité de l'information, laquelle dépend du temps, du lieu (e.g. de la langue) et du contexte, ainsi que des les méta-données pertinentes ;
- Des outils puissants sont nécessaires pour soutenir l'intégration, l'intendance et la gestion de la qualité des données ;

3 D'entropie minimale

4 e.g. service de l'état civil, registre du commerce

5 Données et métadonnées

6 Données et métadonnées

- Les modèles de contrôle d'accès et de journalisation doivent garantir le respect de la réglementation et des règles de l'organisation ;
- La mise en œuvre doit satisfaire les exigences de performance et de sécurité ;
- Enfin et surtout, les processus doivent être mis au point avec la même précision : une organisation sans failles, des ressources suffisantes et une formation appropriée sont nécessaires pour assurer l'intendance des données, la gestion de leur qualité, de leur contrôle, de leur conformité et de leur gouvernance.

5 Prescriptivité douce

Lorsque la diversité et la sophistication des systèmes utilisant les données référentielles augmente, des défis spécifiques surgissent :

- Transactions longues : plus il y aura de sources de données, plus il y aura de conflits à traiter, certains par une intendance manuelle ; il n'est pas possible de garantir que les données de référence seront mises à jour instantanément ; lorsque des tâches s´appuyant sur les données de référence nécessitent qu'une mise à jour soit prise en compte immédiatement, ils doivent bénéficier d'une isolation appropriée des données obsolètes tout au long de leur transaction ;
- Mises à jour différées : selon le métier, il peut être nécessaire d'optimiser la prise en compte de certaines mises à jour, par exemple pour les traiter par lots, ou encore en les suspendant le temps qu'un processus déterminant pour leur cohérence soit terminé. Dans les deux cas, les mises à jour de ces informations doivent être gelées en tant que de besoin pour ces utilisateurs du référentiel ;
- Vérité contextuelle : dans certains cas, bien que correcte, une information référentielle peut être non pertinente pour un sous-ensemble du système d'information ; ce peut être le cas en particulier lorsqu'une information plus précise n'est à partager que de façon restreinte.

Abandonner le principe même de prescriptivité face à ces défis est la pratique la plus courante : des ensembles de données supplémentaires, faiblement intégrés sont créés pour répondre aux besoins respectifs des diverses activités, sacrifiant donc au passage au moins pour partie les bénéfices d'intégration référentielle.

Pourtant, un outillage et une organisation appropriés permettent d'intégrer toute cette complexité sans sacrifier la cohérence. Cette "prescriptivité douce" est construite sur les principes suivants :

- Pour chaque type spécifique d'information, le concept de valeurs de référence canoniques uniques perdure ; pour chaque catégorie de données, une autorité désignée a le dernier mot concernant ces valeurs lorsque des conflits se produisent ; ces valeurs canoniques sont mises en avant : elles sont les valeurs par défaut et restent accessibles dans tous les cas ;
- Des valeurs contextuelles peuvent surcharger les valeurs canoniques pour un périmètre donné lorsque c'est pertinent, soit pour fournir l'isolation nécessaire à une transaction, soit durablement pour quelque raison fonctionnelle ;
- Les requêtes pour mettre à jour les données de référence peuvent spécifier que les données fournies sont délibérément contextuelles, ou bien qu'elles doivent être retenues comme telles au cas où elles ne seraient pas acceptées comme canoniques par le processus d'intégration de données;
- Des métadonnées (statut, commentaires…) permettent de décrire les raisons qui sous-tendent les valeurs contextuelles et de gérer leur cycle de vie ;
- Les données canoniques sont accessibles à tout utilisateur ou processus qui a accès à une surcharge contextuelle correspondante, permettant à tout moment l'analyse des écarts.

La prescriptivité douce fait davantage que donner de la flexibilité processus métiers : elle est un outil irremplaçable de gestion de la qualité des données, grâce à sa capacité de collationner de façon ordonnée des informations qui révèlent :

- Les valeurs effectivement utilisées, potentiellement plus précises ou à jour que celles du référentiel ;
- Des problèmes de modélisation de données, tels que la confusion de concepts ou une erreur de cardinalité ;
- Des dysfonctions dans les processus métiers.

La flexibilité de la prescriptivité douce exige une gestion rigoureuse afin de prévenir les abus. En particulier:

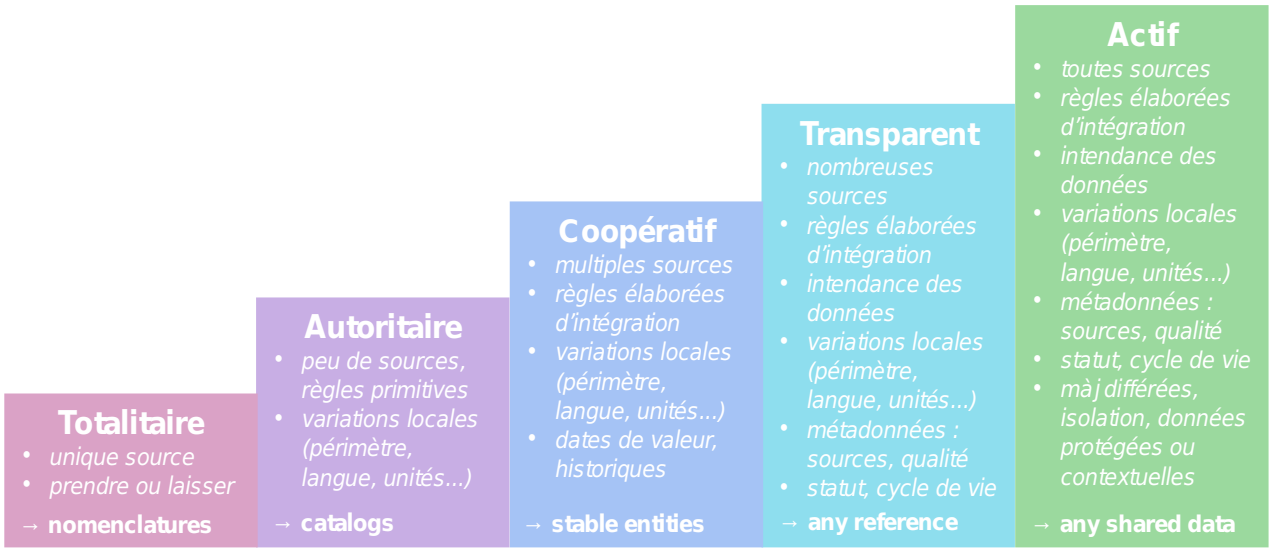
- Les concepts qui sont ontologiquement distincts (par exemple service et centre de coûts, nom légal et nom d'usage...) doivent être identifiés et clairement séparés dans des conteneurs de données distincts, et non implémentés comme des variantes contextuelles l'un de l'autre ;
- Les concepts qui sont similaires à des données référentielles existantes, mais sont en fait spécifiques à une activité, ne doivent pas figurer du tout dans le référentiel.

Des processus qualité doivent surveiller les données référentielles afin d'en assurer le nettoyage et d'identifier les signes révélant des défauts de conception ou des dysfonctionnements organisationnels.

6 Maturité des référentiels

Les modèles habituels de maturité des données référentielles décrivent la maturité de la gouvernance mais manquent à préciser leurs implications techniques concrètes.

Le diagramme ci-dessous, au contraire, est centré sur les réalisations techniques qui, moyennent une organisation rigoureuse, rendront cette gouvernance possible et opérationnelle.



Les limitations technologiques du référentiel de l'organisation peuvent empêcher la gestion des catégories de données les plus exigeantes du fait de leur complexité (sources, variabilité, qualité, utilisation).

En d'autre termes, la capacité technique d'un référentiel détermine le périmètre maximal auquel il peut s'appliquer de façon prescriptive. Les ambitions de la gouvernance des données et le développement technique des référentiels doivent donc être attentivement synchronisés.