

La gouvernance des données

Introduction

La gouvernance des données peut être définie comme un système alliant une bonne connaissance des règles applicables à une définition claire des rôles et responsabilités des acteurs dans tous les processus générant des données ou en lien direct avec celles-ci.

Il s'agit donc de bâtir une stratégie qui est nécessairement transverse à une organisation, publique ou privée. De cette stratégie découleront des méthodes pour réaliser des actions concrètes visant principalement à réexploiter les données. Les problématiques à prendre en compte sont de natures diverses, elles ne se rapportent pas exclusivement aux données en tant que telles. À titre d'exemple, nous pouvons citer les questions de confidentialité, que ce soit pour la protection du secret de la défense nationale ou bien pour préserver le secret médical ou celui des affaires, la propriété intellectuelle, les études de marché pour cibler des publics spécifiques, etc. Plusieurs axes de travail doivent être articulés parmi lesquels la protection et la confidentialité, la qualité de la donnée, la valorisation et l'innovation.

L'objectif de ce dossier est double : il s'agit de montrer les différentes facettes de la gouvernance des données tout en donnant des exemples permettant de comprendre comment les archivistes peuvent se situer dans l'organisation nécessaire à la maîtrise des données.

Coordination :



Sarah Clinet

Responsable du management de l'information
Ministère des Armées



Lourdes Fuentes Hashimoto

Département archives et *records management*
Groupe Total

La donnée pour tous ?

Toutes les organisations portent un intérêt croissant aux données qu'elles produisent. L'objectif premier de cet intérêt nouveau étant la réutilisation de ces données, que ce soit dans le cadre des politiques d'*open data*, de la promotion de la transparence administrative, de mise en œuvre de *business intelligence* – informatique décisionnelle – ou bien à des fins de création de nouveaux services pouvant parfois aboutir à la redéfinition de modèles économiques.

Ces dernières années, la donnée a souvent été présentée comme le nouvel « or noir ». Si la comparaison avec le pétrole peut être discutée, elle a le mérite de mettre en évidence le fait que l'identification de gisements de données est insuffisante, encore faut-il organiser le raffinage et penser ensuite aux traitements, à l'utilisation, voire à la distribution de ces données. Cette première approche de la donnée est à replacer dans le contexte du développement des plateformes comme Facebook. Nous sommes passés progressivement d'une approche de redistribution telle quelle de la donnée brute à des fins commerciales, avec la vente des jeux de données, à la ré-exploitation et au retraitement de la donnée pour inventer de nouveaux usages. Aucun secteur d'activité n'est épargné : nous avons vu l'émergence d'entreprises tournées entièrement vers l'économie numérique qui placent la donnée au cœur de leur développement ; le marketing prédictif en est un exemple concret. Les institutions publiques n'ont pas résisté non plus à l'appel de « la donnée pour tous ». L'État français s'est ainsi saisi de la problématique en mettant en place Etalab, une structure dédiée à l'élaboration et à la promotion d'un État plateforme. Nous entendons désormais parler de la *data science*, assortie des promesses de l'intelligence artificielle. En parallèle, le terme d'algorithme n'a jamais autant fait parler de lui. Les administrations et les entreprises mettent en place des missions internes visant à identifier les données et à en définir les usages possibles par le biais des *data officer*. Dans ce contexte, d'autres préoccupations – et d'autres enjeux – sont apparus : le respect de la vie privée et la protection des données personnelles, ou encore la peur de la surveillance de masse, l'encadrement des transferts des données, l'essor de la cybersécurité, etc. La donnée pour tous ? Oui, mais avec des garde-fous. Il ne fait aucun doute que l'intérêt pour la donnée n'est pas une tendance passagère : il structure pleinement l'activité économique mondiale, comme en témoigne l'essor des GAFA (Google, Amazon, Facebook, Apple) et toutes les politiques de *data innovation* essayant de reproduire le succès de la Silicon Valley qui demeure un véritable Eldorado, non seulement pour de nombreux États mais aussi pour les nouvelles générations arrivant sur les différents marchés de travail. De fait, la donnée a envahi nos vies quotidiennes : on entend parler tous les jours de vols et de fuites de données, de leur vente illicite, de la naissance ou de l'échec d'entreprises tournées vers l'exploitation de la donnée à des fins diverses. La donnée semble d'autant plus à portée de main que l'on a vu se développer des outils permettant de les manipuler très facilement ; nul besoin d'une formation d'ingénieur avec les outils d'exploration et de visualisation des données, par exemple.

La donnée pour tous ? Oui, mais. Pour tirer tous les bénéfices de la donnée, il y a un important travail à réaliser en amont : identifier les ensembles de données, les qualifier, analyser leur qualité, les exporter, les traiter, les modifier, les manipuler, définir leur cycle de vie, comprendre et appliquer la réglementation afférente en fonction du secteur d'activité, mettre en place des garde-fous quelle que soit l'échelle de traitement des données, conserver la donnée dans la durée et préserver son caractère opposable le cas échéant. Autant de questions qu'il est impossible de traiter sans une organisation adéquate faisant intervenir de très nombreux métiers. La gouvernance des données est un travail collectif, la donnée par tous avant d'être la donnée pour tous.

Comprendre comment sont produites les données revient à s'intéresser aux personnes qui les produisent et au cadre de leur activité. Car la source de la donnée est bien l'humain. Le *records management* n'a jamais été aussi nécessaire, les compétences des archivistes n'ont jamais été aussi pertinentes. En parallèle, cet aspect fondamental du métier de l'archiviste qu'est l'évaluation n'a jamais autant dû être réinterrogé. Ainsi, et jusqu'à maintenant, définir la durée de conservation d'une information revenait à s'interroger sur le temps pendant lequel cette dernière serait utile pour le métier qui l'avait produite, aurait valeur de preuve pour l'organisation et serait opposable dans le cadre d'un éventuel contentieux. À l'issue de cette durée, certaines données, jugées désormais inutiles, étaient éliminées, quand d'autres étaient conservées définitivement à titre patrimonial. Alors que se généralise cette seconde vie des données, qu'on les ré-exploite à des fins de transparence, de pilotage ou de création de nouveaux services, il est nécessaire que le *records manager* sache accompagner cette mutation de leurs usages.



Sarah Clinet

Responsable du management de l'information
Ministère des Armées



Lourdes Fuentes Hashimoto

Département archives et *records management*
Groupe Total

La qualité des données



© Diego PH on Unsplash

Pourquoi faire de la qualité des données ?

La première des raisons, qui peut sembler futile au premier regard, est que la mise en qualité des données est obligatoire, car imposée par le réglementaire. Les exemples de conformité abondent : traçabilité des aliments et des médicaments, normes financières IFRS¹, contribution à la lutte contre le blanchiment d'argent, respect des exigences CNIL ou RGPD, étiquetage précis des denrées alimentaires, etc.

Mais, de manière plus générale, nous vivons dans un monde où l'information pertinente a d'autant plus de valeur que les volumes de données explosent d'année en année. Qu'un lot de lait infantile n'affiche aucune date de péremption ? Serait-il acceptable qu'un hôpital vous prélève du sang à chaque visite afin de déterminer votre groupe sanguin, encore et encore ?

Toutes ces questions que nous nous posons en tant que citoyens, consommateurs ou professionnels interrogent de la même manière les organisations publiques ou privées. L'assurance maladie peut-elle se permettre de confondre deux bénéficiaires ? Serait-il acceptable que des versements faits sur un compte en banque améliorent drastiquement les comptes d'un tiers ?

S'il est de la responsabilité d'un individu de – ou tenter de – contrôler la qualité de ses données personnelles, imaginez la complexité d'un chantier de mise en qualité des données relatives à plusieurs millions de détenteurs de comptes pour une banque nationale ou du fichier client d'une grande enseigne de la distribution.

1. Les IFRS (*International financial reporting standards*) sont depuis 2005 les normes comptables internationales. Elles ont remplacé les normes IAS (*International accounting standards*).

Gouverner les données avant qu'elles ne nous gouvernent

Plus abrupt encore est le fait que sans mise en qualité des données, nous serions condamnés à vivre sous le régime de la terreur : celui de la non-qualité. Ce qui est déjà difficile à conduire au niveau d'un individu devient ingérable au niveau d'une organisation si l'on ne met pas en place un dispositif dédié à la mise en qualité des données.

Voyons quelques cas et risques classiques auxquels vous serez confrontés en cas d'absence d'un dispositif dédié à la mise en qualité des données.

• La sémantique des données n'est pas partagée

Autrement dit, la signification d'une donnée, le sens qu'on lui donne, n'est pas unifiée au sein de l'organisation. Par exemple : dans le cadre d'un groupe bancaire aux nombreuses filiales, les éléments composants la rémunération variable des collaborateurs ne sont pas univoques. Si ceci ne pose aucun problème visible au niveau de la filiale, la holding se retrouve incapable de consolider efficacement la masse salariale et de mener des études sur l'efficacité des politiques de rémunération variables.

• La vision d'un même objet n'est pas unifiée

Dans la plupart des entreprises, la représentation d'un même objet référentiel n'est pas unifiée et, comme un prisme, est très orientée par l'activité même du service. Par exemple, comment se représente la structure organisationnelle de l'entreprise pour différents services de la même entreprise ? Pour la direction commerciale de la distribution, il s'agit du maillage des magasins (zone/pays/région) ; pour la direction financière, d'une consolidation de centre

de coûts et de profits ; pour la direction juridique, d'un assemblage de sociétés constituant le groupe ; pour la direction des ressources humaines, d'un assemblage de relations hiérarchiques ou de relations fonctionnelles.

• **Les métadonnées associées aux données ne permettent pas :**

- d'exécuter normalement les processus métiers : le producteur des « données maîtres » n'est pas clairement identifié, les règles de gestion ou de calcul ne sont pas décrites, les flux ou processus sollicitant les données ne sont pas clairement identifiés, le cycle de vie des données ne peut pas être mis en œuvre, etc. ;
- de réaliser une analyse qualitative des données : pertinence, confiance, précision, fraîcheur et accessibilité des données, non-conformité aux référentiels, etc.

• La mise en cohérence des données est contestée par des directions souhaitant conserver une position dominante sur leur « périmètre informationnel ».

• Les projets de systèmes d'information ou de transformation n'incorporent pas suffisamment tôt les chantiers relatifs aux données.

• Et malheureusement bien d'autres cas.

Les ressources nécessaires à un projet de mise en qualité des données comprennent inmanquablement : un sponsor au bon niveau ; un dispositif de gouvernance déployant des processus compréhensibles ; des équipes convaincues et reconnues.

Un sponsor au bon niveau, pas moins

Le sponsor d'un projet de mise en qualité des données est à coup sûr le point le plus important dans la réussite de ce dernier. Par « au bon niveau », l'on doit comprendre « au bon niveau de décision » pour pouvoir acter des orientations, valider les plans d'action, capter les ressources nécessaires à la mise en œuvre du projet. Mais « au bon niveau de décision » n'est pas suffisant. Le sponsor doit en outre porter une conviction : celle de la nécessité impérieuse du projet.

La mise en qualité des données est en effet un chantier permanent : nettoyer les écuries d'Augias est une chose, les maintenir propres en est une autre. Le sponsor doit avoir à cœur de soutenir l'énergie des équipes lors des phases difficiles, de faire entrer les activités liées aux données dans le quotidien et de ne jamais céder sur l'impérieuse nécessité de mener les travaux à terme.

Si un sponsor au bon niveau de décision, impliqué, sachant et porteur des enjeux du projet est un avantage indéniable, il ne peut à lui seul déployer l'ambition d'un tel chantier. Si le sponsor donne l'impulsion initiale, il est nécessaire de mettre en place un dispositif pour ancrer la réalité de la qualité des données dans le quotidien de l'organisation.

Un dispositif de gouvernance adapté déployant des processus compréhensibles

Se poser la question de la gouvernance des données revient à se demander : « Qui fait quoi à quel moment afin d'atteindre quel résultat ? De quoi avons-nous besoin maintenant et à l'avenir, en termes de données, pour soutenir les activités métiers ? »

La première question relève autant de l'ensemble de la structure que des individus. Du niveau stratégique jusqu'aux opérationnels, chacun doit savoir ce qu'on attend de lui, et ce dans quel laps de temps ou selon quel agenda. Le résultat visé doit lui aussi être communiqué clairement. En effet, la valeur générée par les opérations de saisie, de modification ou de correction des données peut parfois échapper aux responsables d'information chargés de leur maintenance. Sans partage de cette finalité, les tâches liées aux opérations apparaîtront vite rébarbatives, pour ne pas dire vides de sens. Un exemple ? Dans le

cadre d'un projet d'intégration d'un progiciel au sein d'une enseigne de l'habillement, un directeur des collections s'était insurgé contre la charge administrative induite par la saisie d'informations inutiles pour ses équipes. De son point de vue, nulle valeur n'était créée par l'intégration de ces données. Et pourtant, ces données trouvaient leur résurgence sur l'étiquette embarquée sur chaque vêtement, sans laquelle aucun article n'aurait été commercialisable en magasin. Faire visiter aux équipes des collections les autres services de l'entreprise a rapidement corrigé cet écueil.

La seconde question qui relève de la gouvernance est de s'assurer que les données permettent de soutenir les processus actuels, mais aussi et surtout de répondre aux besoins en données nécessités par les futurs objectifs stratégiques de la structure.

Des équipes convaincues de l'importance de la mise en qualité des données

Comme l'enseigne la roue de Deming (ou le supplice de Sisyphe), la mise en qualité des données est un chantier permanent. En cas d'interruption de cette politique, la qualité des données baisse drastiquement et les processus en pâtissent immédiatement : les temps passés en correction *a posteriori* augmentent, la confiance dans l'information manipulée baisse, les contrôles manuels se généralisent et l'efficacité des collaborateurs diminue significativement.

La qualité des données doit être l'affaire de tous et devenir l'une des composantes intrinsèques à tous les postes de l'entreprise, tant au niveau de la création que du contrôle. La notion « d'entreprise étendue » permet même d'élargir le raisonnement aux contributions des fournisseurs et des clients. Ces derniers étant tantôt producteurs ou consommateurs de données, ils ont d'évidence aussi leur mot à dire sur la qualité des données qui les concernent. Et gare à celui qui ne prendrait pas en compte leurs demandes de modifications !

L'humain, comme bien souvent, est au centre de tout. De son engagement dépend la réussite de ce vaste chantier qu'est la mise en qualité des données.

Et l'archiviste dans tout cela ?

Une application efficiente du cycle de vie du produit est un avantage concurrentiel pour un bon nombre d'industries de pointe. Ainsi, au sein des firmes aéronautiques, les données doivent être maintenues jusqu'à soixante-dix ans entre le lancement d'un projet de R&D visant la conception d'un nouveau modèle d'avion et son retrait de service. Ces durées peuvent même s'allonger afin de contribuer à l'analyse du risque aérien ou pour comprendre la nature de certains accidents. Or, ces délais d'engagement dépassent bien souvent la capacité de projection des organisations. Mis à part l'archiviste, qui aujourd'hui est capable d'organiser la mise en qualité des données à un horizon aussi lointain, là où les objectifs habituels de pilotage ne voient guère plus loin que le trimestre ou la fin de l'exercice comptable ?

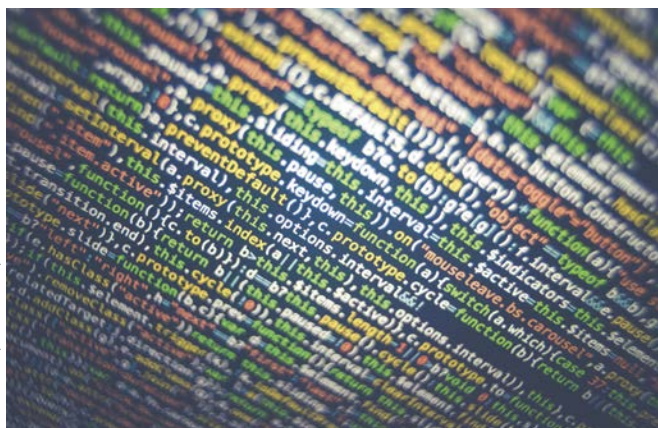
Tout comme le sponsor, la gouvernance et des équipes convaincues, l'archiviste est un atout indéniable de toute politique de mise en qualité des données, à la fois par son expertise dans leur gestion mais aussi par cette capacité de projection vers un « long terme » faisant souvent défaut aux autres fonctions de l'entreprise.



Jean-Yves Mahe

Responsable gouvernance SI
Ministère des Armées

Innover par la donnée



© Markus Spiske on Unsplash

Les données sont partout et sont légion. Si leur utilisation nourrit bien des fantasmes, elle constitue également un formidable levier pour innover, et ainsi créer, améliorer ou transformer des services. Le tout est d'identifier les données et de les utiliser à bon escient.

Une innovation réussie, c'est d'abord une bonne idée, puis une exécution rapide et enfin un fort engagement des utilisateurs. L'innovation apporte une rupture dans l'usage ou dans la perception du service ou de l'image de l'organisation par ses utilisateurs. Toute innovation est contextuelle : elle peut être nouvelle dans un secteur traditionnel et déjà dépassée dans un autre. Enfin, idéalement, lorsqu'elle est digitale, l'innovation n'aurait pas pu être mise en œuvre cinq ans plus tôt. Elle est permise par les évolutions technologiques et répond à des problèmes de société contemporains. Comme l'innovation irrigue tous les champs, qu'ils soient managériaux, financiers ou sociaux, elle implique de travailler en synergie avec toutes les parties prenantes. C'est un processus collaboratif et itératif.

L'une des clés de l'innovation ? Les données : « *In God we trust. Everyone else bring data* » (Nous croyons en Dieu, que tous les autres apportent des données). Mike Bloomberg, entrepreneur et homme politique, explique que cette phrase a guidé toutes ses actions, notamment philanthropiques, avec succès. Pas d'innovation sans données aujourd'hui, c'est le mantra que l'on ne cesse d'entendre en innovation digitale.

Les données que nous avons déjà

La *data*, c'est souvent tout ce qu'on ne sait pas que l'on sait. Les organisations sont assises sur des trésors de données sans vraiment s'en rendre compte. Dans tout processus d'innovation impliquant des données, il s'agit en premier lieu de les valoriser, de les caractériser et de les identifier.

Commencer par une analyse des données statiques est la première étape, en impliquant notamment les acteurs de l'informatique, du contrôle de gestion et du métier. Une fois l'état des lieux établi – ce qui existe –, le travail de fond consiste à établir une base de données unique et fiable, qui doit idéalement être alimentée par de la donnée dynamique, vivante, en temps réel. Un point crucial ici est de s'assurer que les données sont fiables, qu'elles sont exhaustives et que l'interopérabilité est recherchée, c'est-à-dire que les systèmes d'information communiquent entre eux.

Toutes les données sont-elles utiles ? Oui potentiellement : leur grand nombre impose toutefois de faire certains choix prioritaires qui doivent être guidés par un schéma d'ensemble, ce qui suppose d'avoir déjà une idée de ce que l'on recherche. Attention à ne pas minimiser le travail que cela représente.

Par exemple, pour l'organisation du datathon (voir encart) de Paris Habitat en juillet 2018, nous avons commencé par mettre en place un *data lake*, c'est-à-dire un référentiel de stockage avec une grande quantité de données brutes dans leur format natif. Composé de données internes et externes avec une nomenclature minimale, le système est alimenté en continu par un système d'API – une interface de programmation applicative par laquelle un logiciel offre des services à d'autres logiciels – et les données sont catégorisées et accessibles via une clé unique. Avec une architecture à plat (les données ne sont pas classées dans des fichiers ou dossiers), ce référentiel permet au *data scientist* de s'y plonger pour les faire parler, un peu comme un archiviste et sa multitude de documents de nature, volume et pertinence hétérogènes.

Des sources de données... à créer

Au-delà des sources liées à la structure ou à son environnement, il est possible de mettre en place des systèmes collectant des informations. Ainsi, dans le cadre de l'innovation dans les services publics en Seine-Saint-Denis et le SP Lab (Service Public Laboratoire) mis en place en 2016, nous avons lancé, auprès du public de la Maison des personnes handicapées du département, le robot R2D3. Il permet de consulter son dossier en ligne (ce qui implique déjà un accès à des données) ou de donner son avis sur la qualité de l'accueil. R2D3 collecte ainsi des informations sur ce qui est demandé par le public pour améliorer le service, et permettre ainsi l'innovation.

Poussé un peu plus loin, le principe est le même que l'utilisation de l'intelligence artificielle dans le sport de haut niveau : les athlètes s'entraînent dans des « vraies » conditions ou en réalité virtuelle avec des capteurs qui collectent des données précises pour comparer les résultats et se perfectionner. Plus proche de nous, des studios *createch* par exemple proposent des expériences immersives permettant la collecte de données, comme récemment une plateforme web immersive de formation aux règles de sécurité pour la SNCF, présentée au salon Viva Technology par le studio Blumenlab. Il est ainsi possible de créer des expériences et d'en collecter les données pour améliorer les procédés ou la performance.

Peut-on prendre en compte toutes les données ?

Le règlement général sur la protection des données (RGPD) a pris une place très importante dans l'actualité. Appliqué dans l'Union européenne depuis le 25 mai 2018, il est un garde-fou nécessaire pour tout ce qui concerne les données personnelles (ou données à caractère personnel) qui permettent d'identifier une personne physique, directement ou indirectement. Toute structure publique ou privée, et particulièrement l'État et les collectivités territoriales, a une responsabilité concernant l'utilisation de ses données, mais cela n'empêche aucunement le partage avec des tiers, y compris des entreprises. Par exemple, *api.gouv.fr* – une start-up d'État de la communauté *beta.gouv.fr* intégrée à la Direction interministérielle du numérique et du système d'information et de communication (DINSIC) – référence les API de l'État utilisées par les entités publiques et en grande partie par le secteur privé, que ce soit pour obtenir des pièces justificatives pour des personnes physiques ou morales, pour des données de transport ou encore pour géolocaliser une adresse. Quand il s'agit de données à caractère personnel, il n'est possible de les communiquer à des tiers qu'à deux conditions alternatives : soit l'utilisateur a donné son consentement explicite pour chaque réutilisation/transfert, soit les données sont anonymisées sans possibilité de ré-identification. Les règles sont strictes en la matière, ce qui peut parfois compliquer le travail sur les données en amont. Il s'agit ici de faire preuve de la créativité encouragée dans les processus d'innovation tout en respec-

tant le cadre juridique. Par exemple, pour le datathon de Paris Habitat organisé en juillet 2018, il n'a pas été possible de mettre à disposition des données très précises, non pas parce que l'organisme ne les avait pas mais afin de préserver l'anonymat des locataires. Le niveau de granularité a ainsi été aussi faible que possible. Les participant.e.s au datathon ont dû innover avec cette contrainte.

Quelles innovations ?

L'innovation peut-être disruptive – apporter quelque chose de complètement nouveau qui transforme le *business model* ou les moyens de production – ou incrémentale – qui apporte des améliorations ne modifiant pas profondément les modalités de fonctionnement. Le smartphone est l'exemple d'innovation disruptive par excellence. L'essentiel étant qu'il ne s'agit pas d'innover pour innover ou d'expérimenter pour expérimenter... Dit autrement, l'idée n'est pas de se « faire plaisir » avec une innovation qui ne rencontrerait pas son public. Car ce qui fait une innovation réussie, c'est avant tout l'engagement des utilisateurs. Pour les applications qui sont sur les smartphones, par exemple, seulement 15 % du temps passé sur un téléphone portable se fait sur une application téléchargée !

Il s'agit de réfléchir en termes d'usages, d'envisager de nouvelles actions pour offrir un service plus en adéquation aux besoins des usagers ou de rechercher une plus grande efficacité des services avec l'appui des nouvelles technologies, et donc de considérer les problèmes à résoudre pour les utilisateurs finaux avant ceux de la structure qui innove.

Pour le datathon de Paris Habitat de juillet 2018, nous avons défini trois thèmes sur lesquels Paris Habitat souhaitait réfléchir en *open innovation*, avec comme mot d'ordre de ne pas brider l'imagination des participant.e.s. L'analyse de données nous a permis de déduire des comportements et d'identifier les problèmes tels qu'ils sont perçus. Ce travail collaboratif a permis d'identifier collectivement les principaux challenges à relever dans les prochaines années. Pour innover avec les données, il s'agit de « réfléchir avec les mains », en expérimentant, en préférant l'itératif au produit fini, parfait. S'atteler à répondre à des défis, à des problèmes, à des irritants permet de tirer le meilleur des données. Cette analyse poussée des données peut faire ressortir des schémas récurrents, jusque-là cachés (parfois oubliés). En cela, l'innovation est une forme de désobéissance et peut venir perturber une organisation et un plan de travail bien établis. La tentation d'un « solutionnisme technologique » est grande, mais la réflexion, l'apport réel pour les utilisateurs finaux et l'impact de l'innovation doivent prendre le pas sur l'attrait de la nouveauté.



Carole Stromboni

Experte en transformation digitale
et en innovation

Datathon

Aujourd'hui, les données sont partout, sur les réseaux sociaux ou les commerces en ligne, chez les opérateurs télécoms, à chaque passage en caisse, dans les capteurs météo... ou encore dans les systèmes de données métier. Elles prennent de plus en plus de place dans notre quotidien et les récentes avancées technologiques permettent de les analyser avec une puissance de calcul inégalée. C'est pourquoi le format du datathon se généralise. Un datathon, c'est comme un hackathon, généralement dans le même format de deux ou trois jours, consécutifs ou non, mais où le personnage principal est... la donnée. Il s'agit d'une course contre la montre pour imaginer de nouveaux services en rassemblant des personnes de profils divers (*datascientists*, développeurs, salariés, start-up, utilisateurs finaux, étudiants, etc.). Les thèmes peuvent être très larges ou très restreints. Plus le thème est précis, plus les innovations seront disruptives. À l'inverse, plus le thème est large, plus les innovations seront incrémentales.

À partir de la donnée, en reliant les éléments et en les caractérisant, l'innovation prend une nouvelle dimension.

Data lake (lac de données)

Un *data lake* est un référentiel de stockage conservant une grande quantité de données brutes dans leur format natif jusqu'à ce qu'elles soient nécessaires. Il présente une architecture à plat, c'est-à-dire que les données ne sont pas classées dans des fichiers ou dossiers. C'est comme une grande source d'eau à l'état naturel. Le principal défi n'est pas de le créer mais de tirer profit des opportunités qu'il représente. Les structures ayant développé des *data lakes* avec succès les améliorent progressivement, à mesure qu'elles se demandent quelles données et quelles métadonnées sont importantes pour leur activité.

Les caractéristiques d'un *data lake* sont :

- sources de données hétérogènes ;
- alimentation continue via un système d'API ;
- nomenclature et croisement de données : idéalement des données brutes dans leur format natif avec toutefois une nomenclature minimale à l'import (modèle de données et harmonisation des données), afin de pouvoir les exploiter ;
- accessibilité : les données sont catégorisées et accessibles via une clé unique dans le but d'être exploitées facilement pour servir des applications métiers.

Les données, du gisement à la valorisation



© Dennis Kummer on Unsplash

Les données sont considérées comme une mine d'or pour l'entreprise qui en possède. Si la métaphore permet de sensibiliser au potentiel que représentent ces données, elle témoigne également du travail nécessaire pour atteindre cette valorisation supposée. En poussant le parallèle avec le précieux métal, il faut au préalable savoir que ces données existent et savoir où et comment y accéder, qui les possède. Il faut ensuite en explorer le contenu, en comprendre le sens puis le potentiel, s'assurer de leur qualité, imaginer un usage et le réaliser. Sans oublier les autres étapes en amont et en aval, comme la captation des données, l'innovation à partir des données, ou encore l'élaboration du business model afin de les commercialiser, qui ne seront pas ici évoqués.

En 2015, nous avons créé le *datalab* au sein de la direction information, veille et archivage du groupe Total dans l'objectif de valoriser des travaux de veille et de recherche d'information. Ce *datalab* permet de réaliser trois actions : une exploration des données, du *processing* automatisé pour les enrichir et de la visualisation. Rapidement, un fort intérêt s'est manifesté pour ces services, nous poussant à les proposer pour d'autres domaines que la recherche d'information et la veille. Ces trois services, dont l'utilité est détaillée ci-dessous, peuvent également être sollicités séparément selon les projets.

Explorer et comprendre

Explorer des données est l'activité la plus déterminante. Cette exploration produit de fait un diagnostic qui conditionne les actions que l'on va ensuite prendre par rapport aux données en question : corriger le processus de création de ces données, les enrichir, les utiliser pour piloter ou améliorer une activité, les proposer à d'autres équipes ou tout simplement décider de ne pas utiliser ces données.

Les outils que nous utilisons permettent de très rapidement mettre en valeur le potentiel de chaque jeu de données. La connectivité est grande pour injecter un très grand nombre de formats, y compris ceux dédiés aux formes géographiques. De façon automatisée, les adresses deviennent des points sur une carte. Les dates produisent une *time-line*, et les données sont représentées sur de nombreux graphiques, facettes et filtres et permettent de naviguer dans les données et d'en faire ressortir le contenu par le biais de visualisations.

La vision apportée est alors si différente de celle proposée dans une base de données ou dans un fichier Excel que l'utilisateur réalise le potentiel des données utilisées. Il se projette dans un pilotage à partir de ces données et dans la création de nouveaux usages. Par ailleurs, cette exploration des données est toujours le moment d'une découverte de fausses croyances, par exemple sur le mécanisme qui a produit le jeu de données. Cela facilite notamment toute la partie amélioration des processus de création des données en amont.

Enrichir

Les progrès dans les outils utilisés permettent de très facilement croiser des jeux de données pour les enrichir. Par ailleurs, de plus en plus de données sont accessibles, comme c'est le cas sur les portails *open data* dédiés à la réutilisation. L'enrichissement est alors possible à partir d'une multitude d'angles : les référentiels, les adresses, les noms, les valeurs. Tout dénominateur commun permet de rapatrier des données pour compléter un jeu existant.

L'évolution des outils et leur vulgarisation sont ici la clef. Sans être développeur, il est possible d'automatiser la collecte de ces données, leur mise à jour, de compartimenter selon des droits d'accès, d'appliquer une qualification de ces données, d'effectuer l'enrichissement depuis une autre source et d'alimenter des visualisations.

Visualiser

La visualisation est utile à différents niveaux. Lors de l'exploration, elle donne directement un aperçu visuel du contenu des données, et facilite ainsi toute l'exploration définie plus haut. Au-delà de l'exploration, elle facilite deux utilisations des données.

L'utilisateur peut utiliser la visualisation pour retrouver une donnée. De par la facilité de navigation dans les données, l'utilisateur va pouvoir modifier son scénario de recherche et affiner ses critères pour accéder à l'information qu'il cherche. L'autre cas d'usage répandu est l'analyse : pour identifier des faits marquants, des acteurs, des thèmes, des évolutions, et pour effectuer des comparaisons dans le cadre d'une aide à la décision.



Vincent Guigue

Département documentation et veille
Groupe Total

Quand les fonctions ressources font rimer RGPD avec opportunité

Envisager à la bonne échelle et de manière globale les transformations liées au numérique était une évidente nécessité pour le directeur général de la ville d'Échirolles.

Ainsi, un groupe de travail Administration numérique a vu le jour en 2015. Créé dans l'objectif de favoriser une meilleure compréhension des enjeux de la transformation numérique, ce groupe permet une bonne collaboration et des synergies entre services. Dans ce contexte, le service documentation-archives a pu accéder au bon niveau d'information et prendre place dans les discussions avec le directeur général, aux côtés du juridique, de l'informatique et de la communication. Le service a été force de proposition sur la gestion électronique des documents, la dématérialisation des processus, l'archivage électronique, l'*open data*, etc. Cette dynamique a mis en évidence l'interrelation entre les différents projets, et en particulier l'intérêt de réfléchir de manière transverse à la gouvernance des données. C'est donc au sein de ce groupe que la question de la protection des données personnelles a été abordée. Ayant déjà conduit plusieurs projets transverses, le service documentation-archives avait une bonne vision d'ensemble des activités de la ville. Ainsi, le pilotage de la mise en conformité au règlement général de protection des données (RGPD) et le rôle de DPO (*data protection officer*) lui ont été attribués de manière plutôt naturelle.

La mise en conformité au RGPD est une opportunité, puisqu'elle nécessite d'analyser et de gérer les risques, de mettre en conformité les processus internes, de documenter cette conformité. Avec la cartographie des traitements par exemple, le RGPD introduit la mise en œuvre d'une approche qualité pour la gestion de l'information au sein d'organisations où cette approche n'est pas toujours ancrée. Pour la ville d'Échirolles, le travail effectué pour se mettre en conformité est une étape pertinente qui facilite d'autant la préparation à l'archivage électronique. Il sera capitalisable au moment de la mise en œuvre du système d'archivage électronique (SAE) mutualisé, projet mené depuis plus de deux ans avec trois autres communes via un syndicat intercommunal.

Le groupe de travail RGPD formé début 2018 se compose de six agentes de la collectivité : une juriste, la responsable du service secrétariat général, la DSI, la chargée de communication et responsable web, la documentaliste chargée de veille et la responsable du service documentation-archives. Le groupe s'est tout d'abord auto-formé pour acquérir les compétences nécessaires, grâce au MOOC proposé par le Conservatoire national des arts et métiers (CNAM) ou à des journées organisées par le Centre national de la fonction publique territoriale (CNFPT). Il a d'abord œuvré à sa propre structuration, répartition des rôles, planning de la mise en conformité et modalités d'action. Son plan d'action s'est inspiré de celui proposé par la Commission nationale de l'informatique et des libertés (CNIL) :

1. Sensibilisation
2. État des lieux et recensement : inventaires des traitements, quelles données nominatives, service par service
3. Cartographie des données et évaluation des risques, afin de réaliser des fiches action adaptées aux interlocuteurs
4. Plan d'action pour la sécurisation : élaboration d'un document de référence
5. Préconisations, élaboration de procédures
6. Procédure de contrôle de conformité, réponse à l'utilisateur, à la CNIL
7. Bilan et ajustements

Des supports ont été produits pour l'étape sensibilisation : présentation, quiz, guide pratique, modèle de recueil des données personnelles. Les actions se mènent en binôme afin de garantir la qualité du suivi et de maintenir la bonne dynamique du groupe. Les réactions positives et proactives des premiers services rencontrés témoignent de l'intérêt et d'un sentiment de progrès.

Le fait de ne pas faire reposer cette mise en conformité sur une personne recrutée à cet effet est un choix. La collectivité mise sur la connaissance des acteurs et des pratiques dont sont déjà détenteurs les membres du groupe de travail. La dynamique de ce groupe et le choix d'un portage à plusieurs voix laissent envisager de bons résultats.

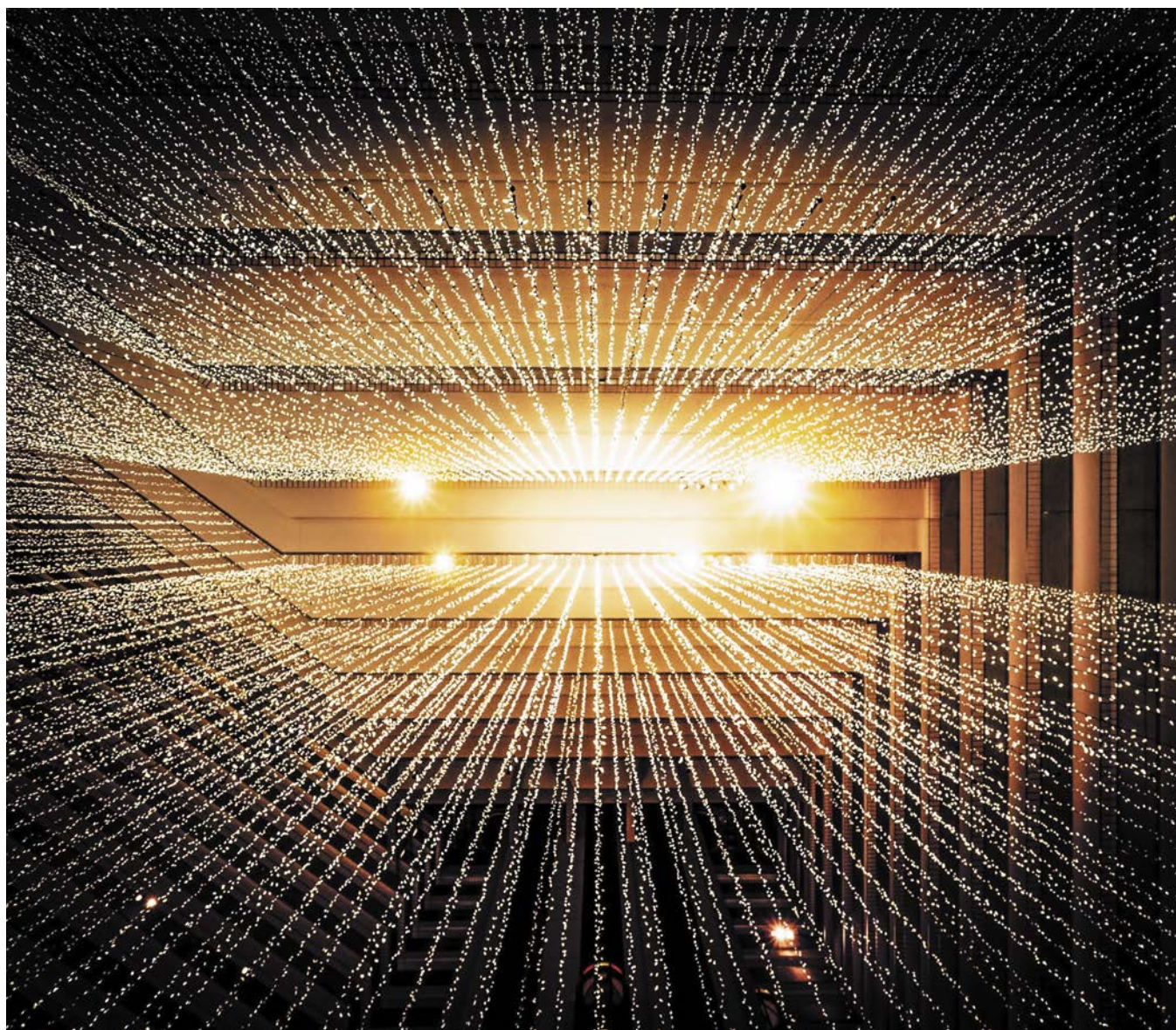
La mise en conformité RGPD, cette entrée « données personnelles » dans les organisations, se fait à marche forcée, mais va permettre une montée en compétences rapide des acteurs. Pour la fonction archives et la fonction documentation, c'est une opportunité de plus saisie pour exercer son rôle dans la structuration de l'information. Dans un environnement toujours papier et numérique, le RGPD donne les rênes à la gouvernance des données.



Pascale Merle

Responsable du service documentation-archives
Ville d'Échirolles

Open data et gouvernance de l'information



© Joshua Sortino on Unsplash

Si, en France, la liberté d'accès à l'information publique peut trouver l'un de ses actes fondateurs avec la Révolution française, l'*open data* tel que nous le connaissons aujourd'hui prend racine dans les pays anglo-saxons, avec notamment l'adoption de l'*open government act* aux États-Unis en 2007.

Le mouvement *open data* s'est focalisé au départ sur la volonté d'obtenir l'ouverture des données publiques, en laissant de côté dans un premier temps les exigences liées notamment à la qualité des données, ne voulant pas en faire un frein à la participation des organisations à ce mouvement. Les politiques plus récentes de l'*open data* montrent désormais, et à plusieurs titres, une volonté de lier ouverture des données et gouvernance de l'information.

Les organisations – administrations, collectivités locales mais aussi entreprises – s'appliquent aujourd'hui à tirer avantages et bénéfices de l'ouverture de leurs données tout en s'astreignant aux obligations qui sont les leurs.

L'ouverture des données : une chance à saisir

L'ouverture des données enclenche nécessairement de nouvelles habitudes de fonctionnement et de travail. Si parfois ces changements peuvent être ressentis par les acteurs comme des contraintes, nous allons voir en quoi ils constituent de réelles opportunités au profit des publics internes et externes aux organisations. C'est en effet en leur faveur que nous procédons à l'extraction de nos données depuis nos systèmes d'information de manière à les leur exposer et à les leur rendre réutilisables.

En interne dans une organisation, un meilleur partage de l'information est indispensable pour casser les logiques de travail en silo, améliorer les performances et au final permettre un meilleur pilotage de l'activité. Par ailleurs, la réutilisation interne des données par d'autres services que leur entité productrice (en anglais *eat your own data*) est, en termes de communication, une bonne manière de valoriser le travail des agents ou collaborateurs. Bien entendu, tout ceci ne peut se faire qu'avec un accompagnement important au changement, l'un

des risques majeurs de non-adhésion des producteurs de données à la démarche étant la peur du jugement extérieur sur leur travail et la qualité de leurs données.

Les données sont souvent appelées « or noir du XXI^e siècle ». On peut discuter la pertinence de cette expression, mais il faut tout même reconnaître que la création de richesse est un objectif important mis en avant par les politiques d'ouverture des données. L'argument du développement économique se retrouve de façon prégnante au niveau des collectivités territoriales, à la recherche de nouveaux ressorts de développement. L'*open data* est vu comme un potentiel d'innovation et de création de nouveaux services : pensons notamment aux secteurs du tourisme, des transports ou de l'énergie. Les administrations comme les entreprises peuvent ainsi faire éclore un nouvel écosystème à partir d'initiatives privées, publiques ou citoyennes.

Comme mentionné plus haut, le secteur public est aujourd'hui soumis à certaines obligations en matière d'ouverture des données, la loi pour une République numérique instaurant ce que l'on pourrait appeler un « *open data* par défaut ». L'ouverture devient la norme, même si celle-ci reste soumise au respect d'autres réglementations (Code du patrimoine, respect de la propriété intellectuelle, industrielle et commerciale, RGPD, etc.). Pour autant, les administrations et les collectivités sont encore loin de répondre à toutes leurs obligations. Certaines organisations publiques font le choix d'une politique d'ouverture ambitieuse qui leur apporte alors un gain d'image en termes de transparence, mais aussi de modernité administrative. En permettant un *empowerment* des citoyens ou des clients, les organisations peuvent satisfaire à l'importante demande de transparence que nous connaissons aujourd'hui, qu'il s'agisse de contrôles démocratiques (choix budgétaires, attribution de subventions, procédures de marchés publics, etc.), mais aussi d'aspects plus consuméristes tels que l'*Open food facts*¹.

On voit donc que les producteurs de données peuvent tirer des avantages divers à mettre en place une politique d'ouverture de leurs données, cette politique, pour être efficace, devant prendre place dans une démarche plus globale de gouvernance de l'information.

Connaître sa production pour l'ouvrir

Si on a pu avoir au départ une vision très technophile et idéaliste de l'ouverture des données, en pensant que cette politique était facile à mettre en œuvre et se limitait à la création d'outils informatiques, il est désormais admis qu'extraire et exposer au public des données ne peut se faire que dans le cadre d'une nouvelle organisation du travail. Toute démarche d'*open data* doit nécessairement passer par une première phase de cartographie de la production informationnelle de l'organisme. Avoir une bonne connaissance des données détenues et produites est indispensable pour étudier ensuite ce qui peut être ouvert ou non, et sous quelles conditions. Ainsi, on constate parfois, par facilité, l'exclusion d'office des données à caractère personnel du champ des données à ouvrir, alors que des méthodes de traitement comme la pseudonymisation peuvent permettre leur mise à disposition si leur contexte d'utilisation le justifie.

Par ailleurs, un apport important de la gouvernance de l'information, sans doute encore peu identifié par les acteurs de l'*open data*, est la prise en compte et la description du contexte de production. Ces éléments de contexte, qui doivent accompagner les données sous forme de métadonnées, sont indispensables à la bonne compréhension et réutilisation des données. Prenons l'exemple des jeux de données dont la composition se base sur des nomenclatures non fournies avec ces derniers : en ce cas, ils deviennent alors une liste de caractères inutilisables.

De plus, de manière plus générale, toute démarche d'*open data* ne saurait contrevenir à la protection des intérêts publics, et des libertés collectives et individuelles que l'État entend protéger au travers notamment du Code du patrimoine. Ainsi, l'*open data* concerne aussi des entreprises ou des délégataires publics, produisant donc des données publiques mais aussi à caractère industriel et commercial. Leur ouverture ne peut donc être envisagée par défaut, mais uniquement après une analyse juridique des risques encourus et du respect des délais de communicabilité.

La gouvernance de l'information apportera aussi une plus forte robustesse à l'*open data*. En effet, si nous voulons permettre aux réutilisateurs de créer de nouvelles richesses à partir de nos données, il faut leur apporter un certain nombre de garanties : origine des données, authenticité, fraîcheur, réutilisabilité technique, etc. Sans cela, le développement de nouvelles activités ou services sera, sinon impossible, bien trop fragile et non pérenne.

Enfin, l'obligation d'ouverture apporte de nouveaux arguments aux gestionnaires de données. En effet, le dialogue avec les éditeurs de solutions informatiques n'est pas toujours simple lorsqu'il s'agit de négocier l'extraction de nos données et métadonnées de l'outil concerné. Engager la négociation via l'angle juridique des obligations auxquelles sont soumises les instances publiques a déjà porté ses fruits. Nous pouvons donc dire qu'une politique de gestion des données apportera à une démarche d'*open data* un gain qualitatif, celui-ci étant nécessaire si l'on veut réellement obtenir de la démarche les bénéfices attendus. Cette politique globale apportera aussi une sécurisation, notamment juridique, aux décideurs et producteurs parfois peu prompts à libérer les données produites. Il faut donc voir l'*open data* comme une déclinaison particulière d'un projet général de gestion de l'information.



Julien Benedetti

Archiviste

1. L'*Open food facts* est une base de données sur les produits alimentaires faite par tout le monde, pour tout le monde. Elle vous permet de faire des choix plus informés, et comme les données sont ouvertes (*open data*), tout le monde peut les utiliser pour tout usage (<https://fr.openfoodfacts.org/>).

Conclusion

Tous, individus, citoyens, collaborateurs, nous produisons des données par de multiples gestes quotidiens, tout aussi insignifiants qu'ils puissent parfois nous paraître, chez nous, dans la rue, au travail, lors de nos loisirs, depuis nos ordinateurs, nos téléphones, avec nos réfrigérateurs, en utilisant notre carte de transports en commun. Ces données personnelles, professionnelles, collaboratives, redondantes, sans que ces qualificatifs ne s'excluent l'un l'autre, sont source par leur réexploitation de valeur ajoutée et d'innovation. Leurs enjeux et leur sensibilité, encore accrue par leur foisonnement, nécessitent qu'on liste ces données, qu'on les organise, qu'on les gère. Ce « on », c'est nous, et « nous » c'est tous les producteurs de données qui, encore moins qu'avant, ne doivent ni ne peuvent se passer d'être individuellement impliqués dans leur gouvernance.

La donnée pour tous ? Oui. Mais, et l'on doit autant s'en préoccuper que s'en féliciter, c'est aussi la donnée par tous, et partout.

Sarah Clinet et Lourdes Fuentes Hashimoto