A N D S I

Association Nationale des Directeurs des Systèmes d'Information

www.andsi.fr

L'approche Open Data et les perspectives de Chief Data Officer

Compte rendu de la présentation du 9 septembre 2014 au Procope

Compte rendu rédigé par Isabelle MAURANGES & ANDSI

En bref…

Henri Verdier, responsable du service du 1er ministre en charge des stratégies d'Open Data défend avec vigueur la donnée. Depuis la Loi Cada, tout citoyen peut consulter toute donnée administrative … L'an dernier Henri Verdier s'est battu et a pu enfin connaître la « réserve parlementaire », la Base de Données des prix de l'essence et la population carcérale … Au-delà de la transparence de l'information, nous voyons naître la notion de « contribution éclairée à la décision publique » et au travers de l'Open Data, l'Etat travaille également l'efficacité de son action publique. L'Open Data est l'antichambre des politiques de Data Science…
Henri Verdier attend maintenant la promulgation du décret de création du Chief Data Officer de l'Etat français, baptisé Administrateur Général des Données. Il connaîtra les données, en validera la qualité au travers d'un rapport annuel et simplifiera l'accès aux données de l'Etat.

Exposé de M. Henri VERDIER

Etalab

Je vais vous parler de stratégie d'innovation, à la frontière des SI, mais la peinture n'est pas encore sèche, il faut tester les hypothèses … Entrepreneur en numérique depuis bientôt 20 ans avant l'internet Grand Public, je faisais des Cédéroms avec Odile Jacob. Avec Georges Charpak, Nobel de physique, nous avons créé « La Main à la Pâte » et fabriqué des outils qui ont équipé un bon tiers des écoles primaires. Ce fut un demi succès car nous n'avons pas trouvé de démarche commerciale magique pour satisfaire 55 000 écoles primaires à tout petit budget.

Voilà dix ans j'étais sur la frontière de l'innovation, Directeur de l'Innovation chez Lagardère Active puis Directeur de la Prospective à l'Institut Bouygues Télécom. Avec beaucoup d'amis, nous avons lancé une belle aventure de Pôle de Compétitivité en créant Cap Digital que j'ai présidé durant huit ans. Nous avons rencontré beaucoup de personnes et vécu de belles d'histoires car nous étions 800 PME, 25 grands groupes et 250 laboratoires de recherche publique et privée. Puis, j'ai écrit « L'âge de la multitude », il y a 2 ans. Je voulais théoriser sur les raisons du succès de l'économie numérique en remarquant que beaucoup de Google, Facebook, Amazon, et autre Twitter s'organisaient pour prendre la création de valeur auprès des clients, de la multitude, de vous et de moi en utilisant des traces d'utilisation, des contributions libres ou des stratégies de plateformes via des API. Il y a trois ans et demi, j'ai fondé une entreprise de mathématiques appliquées au Big Data. C'était le balbutiement du Big Data et des mathématiques avec le Big Data. Je m'étais associé avec des personnes qui font de la finance et des mathématiques appliquées ainsi que d'une médaille Fields. Ma société a été ensuite rachetée par Havas et tourne aujourd'hui avec 35 salariés. Mais le plus important c'est que j'ai découvert qu'on n'avait besoin ni de mathématiques appliquées ni de médaille Fields pour chercher des corrélations et « screener » des millions de lignes en tableur. La règle de trois c'est déjà génial ! Enfin, voici 18 mois, j'ai pris la direction du Service du Premier Ministre en charge du partage et de l'ouverture des données publiques. Je m'occupe d'Open Data et de fil en aiguille je retrouve l'Open Gouvernance car, à partager les données autant voir si des gens s'en servent, si on peut travailler avec elles, les associer à la décision. Nous sommes actuellement en train de créer une fonction de Chief Data Officer pour les services de l'Etat.

Pour ce soir, j'ai prévu une première partie sur « La révolution de la donnée ». En tant que DSI, vous savez que dans le monde des hackers et des geek il existe un slogan « Data is a new code and code is a commodity ». Serait-ce en train de devenir vrai ? N'est-ce pas la donnée qui compte ?

Lorsque j'étais à MFG LAB, les développeurs cherchaient systématiquement dans Google les fonctions pour les « copier/coller » ! Le rapport au code avait énormément changé en 15 ans ! J'ai vu des vagues dans l'évolution numérique : en 2005 c'étaient les grandes stratégies de portail qui comptaient, comme Yahoo, puis ce furent les stratégies de plateformes. Aujourd'hui, le lieu le plus intense en matière de la révolution numérique, c'est la donnée. Et pour cela, il existe nombre de raisons, car on produit de plus en plus de données. Les technologies de capteurs changent la face du monde car des choses qui auraient coûté des fortunes à produire sont entrain de baisser grâce à de petits devices ou des capteurs à moins de 10 centimes d'euro. Pourquoi ne pas « data-ifier » ces phénomènes que l'on n'aurait pas approchés voici 10 ans ? Il y a 5/6 ans à Berkeley, des étudiants ont construit un système d'alerte sismique réseau à base de Smartphone posés sur des tables de nuits. Et ça marchait ! Aujourd'hui, on travaille sur des tissus intelligents captant température du corps, pression, vitesse de déplacement. Dans moins de dix ans, ce sera intégré dans des tissus Grand Public, des chaises, des tables, des autoroutes …

Au travers du Web social, Facebook, Twitter, certains croient à une folie, d'autres remarquent qu'on construit une image sociale de soi pour exister au monde. Mais les clients, nos administrés, partagent tellement de savoir que cela prolifère ! De plus si on croise les technologies du Cloud et les Framework de Big Data, cela devient vraiment peu cher de manipuler des données. En 2009, avec des PC de Surcouf à 2 000€ et 30 millions de clés prêtées par Sybird, j'ai pris conscience qu'une bande de *va-nu-pieds* pouvait manipuler des réseaux de 3 millions de nœuds ! La baisse de coût est majeure, donc nombreux sont les joueurs qui peuvent agir. Il est important de voir que les Framework ont été modifiés pour manipuler du massivement parallèle, des Big Data en flux. Enfin des stratégies sont fondées sur la donnée, et l'utilisent en quantité. L'univers de l'entreprise découvre l'analyse de traces (retargeting publicitaire), le placement publicitaire dans le marketing, les stratégies de plateforme qui modifie son infrastructure pour qu'elle soit le terreau où vont se brancher des applications.
Un des meilleurs exemples de la stratégie de plateforme, c'est le Smartphone : Apple et Steve Jobs voulaient tout faire tout seul et le Conseil d'Administration a contesté ce choix. Résultat : aujourd'hui le device est devenu une plateforme qui a reçu plus de 80 000 applications dans l'Appstore soit plus d'un demi million de développeurs qui contribuent sans être payés ! Et Apple prend 30% du chiffre d'affaires alors que tous ces développeurs sont ravis de pouvoir être là. Pour Google Maps le système est analogue : chez CAP nous avons financé des dizaines de petites entreprises que je désespérais de voir se développer sur Google Maps… Ils sont comme les métayers qui ne possèdent pas leur terre et se mettent en danger … Pour qu'une plateforme réussisse avec plusieurs centaines d'applications, il faut démarrer avec une application phare qui témoigne de la robustesse de la solution. Finalement, je pense que c'est toujours le même métier (stratégie puis coding …) mais le climat est un peu nouveau et la nouveauté, c'est la donnée temps réel.

Aujourd'hui, je m'occupe pour le gouvernement de stratégies d'Open Data (utilisation nouvelle de la donnée). L'Etat travaille sur l'intérêt général qui promeut la transparence. Pour un grand groupe, la stratégie Open Data sera surement différente de celle de l'Etat. Ainsi :

- ouverture de domaines en mettant des fichiers Excel à disposition pour que chacun fasse ce qu'il veut avec ;
- utilisation d'Open API ; les données sont conservées mais quiconque les utilise ;
- Facebook : les requêtes sont autorisées, mais pas l'importation de graphes.

Lorsque vos ressources sont gratuites c'est que la maintenance des données coûte. Ainsi vous irez vers des collectivités « share alike » ou vous remettrez le fruit de votre travail au pot commun.

Dans les stratégies étatiques, on rencontre deux philosophies un peu contradictoires : en France nous avons une tradition complexe de **transparence de l'action publique** s'appuyant sur le préambule de la déclaration des droits de l'homme. De par la Constitution, si vous voulez savoir si vous avez le droit, il faut demander. Rappelons que c'est Lucien Bonaparte qui fonda un service de statistique publique et que les premiers rapports de la Cour des Comptes datent de 1850 ! L'INSEE fut créé après la guerre et prit son statut actuel en 1951, puis arriva la Loi Cada (Commission d'Accès aux Documents Administratifs) qui autorise tout citoyen à réclamer tout document administratif sauf si cela viole la vie privée, la sécurité nationale, si il est couvert par les secrets légaux, ou si il contient de la propriété intellectuelle. Au-delà de la transparence, l'Open Data ouvre la porte aux phénomènes de pollution, à toute la problématique d'anonymisation mais, ceci dit il existe une zone ample où il est possible de travailler sans difficulté. L'an dernier nous nous sommes battus avec succès pour connaître la « réserve parlementaire », la Base de Données des prix de l'essence, de la population carcérale …

Après l'effet « transparence », s'est exprimé le Web 2.0. Certains penseurs américains, fascinés par la puissance à disposition, se sont dit : « Et si les gens aidaient l'Etat comme ils le font pour Wikipédia ? » Le président Obama a vraiment pris la balle au bond : le 21 janvier 2008, jour de son investiture, son premier décret a été le « Transparency Act » pour accélérer la politique d'Open Data. L'Open Data n'est pas que de la transparence (donner une information en page 1448 du PDF de rapport de la Cour des Comptes, est d'une nature différente que de mettre toute l'information à disposition !). Certains citoyens n'apprécient pas quand on leur livre un PDF qu'ils doivent intégralement démembrer puis remettre en forme ! Ce n'est ni une mode, ni un effet Obama, c'est le levier d'une

politique. Open Data ne suffit pas, mais c'est le point d'entrée et ce ne sont pas les mêmes personnes qui comprennent les trois volets concernés : démocratie, innovation et stimulation économique, et, enfin, efficacité de l'action publique.

Il y a 1 an lorsque l'on crée la Haute Autorité sur la Transparence de la vie publique, on se demandait à quoi cela pourrait bien servir … Il n''empêche que c'est par elle qu'on a coincé M. Thévenoud car les vrais fraudeurs ne savent pas forcément maquiller leurs efforts … Au-delà de la transparence, on forge une nouvelle notion : la contribution éclairée à la décision publique car la démocratie participative ne marche pas bien à la façon Ségolène Royale quand tout le monde donne son avis : on n'obtient jamais la meilleure solution. Concrètement, il faut construire les éléments indispensables à la « contribution éclairée » ce qui est un très vaste programme qui est une partie de l'Open Data.

Parlons également de l'innovation : on dit que l'Open Data c'est pour les Start up de 5 à 10 salariés qui développent l'offre culturelle, l'offre immobilière, les transports pour Personnes à Mobilité Réduite … Mais cela apporte également un efficacité stratégique pour les assurances et si l'on veut que cela fonctionne, l'Etat doit alimenter le système.
La troisième forme de création de valeur c'est la donnée qui devient la véritable infrastructure économique critique : On a fait naitre une filière industrielle essentielle. On vient d'essayer de créer le nôtre mais les russes se sont trompés de 8 000 km sur l'orbite géostationnaire… Je pense personnellement que dans les données d'énergie, de transport massif, de santé, il y a de quoi faire naître de nouvelles filières industrielles. Mais il faut avant tout convaincre les détenteurs de données …
Enfin, parlons d'efficacité de l'action publique : quand on oblige les gens à partager leurs données, ils renaclent, mais très vite les barrières tombent et on découvre que le voisin utilisait la même donnée ou mieux, celle qui complémente la mienne. Anecdote : la Ministre du logement n'a pas la base de données des prix de l'immobilier qui existe pourtant chez les notaires et à la DGFIP ! De même, durant des années le Ministère de l'environnement mesurait la pollution des rivières, et le Ministère de l'agriculture celle des nappes phréatiques, mais aucun ne croisait ses bases ! Il existe beaucoup d'endroits où la non fluidité de la donnée crée des dysfonctionnements, des déperditions d'énergie, voire de valeur, ainsi que des méfiances et des rivalités.
L'Open Data doit contraindre un peu tout le monde pour que finalement tous y viennent. C'est un peu comme dans la théorie des Jeux : tout le monde reste sur son optimum local sans partager sa donnée. Quand c'est bien travaillé, quand ça décloisonne, cela fait du bien. Je suis fasciné quand on a des référentiels à partager : en urbanisme Montréal a travaillé une carte urbaine en déclarant un an à l'avance les travaux de voirie, les accidents constatés, les nids de poules … Tout le monde s'est synchronisé et l'on a moins besoin d'autorité pour interdire la circulation car tout le monde se cadre seul … Gain de réunions, gain d'arbitrages …

Troisième raison d'efficacité de l'action publique : quand vous faites l'effort de partager des données vous vous rendez compte que les gens savent les utiliser. Dans les ministères où vous êtes dans un face à face séculaire avec Suez et Veolia, vous avez envie de voir des têtes nouvelles, des stratégies nouvelles, des idées nouvelles. Partout il existe des gens brillants sur votre métier : pensez à ce jeune de 20 ans (prix Data Connexion) qui a construit dans son salon à partir de PC recyclés un ordinateur aussi puissant que celui dont disposait la météo en 2007 et qui alimente 15 jours de météo pour les amateurs de parapente … Je crois passionnément à toute la famille des data-driven stratégiques car sans données on ne peut rien. L'Open Data c'est l'antichambre des politiques de Data Science ...

L'Open Data change, nos stratégies de développement : Lorsque j'ai commencé chez ETALAB, le site ne marchait pas bien, et coûtait très cher (1 million d'euros par an pour un portail dont 400 K€ de sécurité… les données étant ouvertes). Il fallait parfois 8 jours pour partager des fichiers et il m'a fallu un mois pour comprendre l'indécence de la filière de contrôle a priori … Il faut assez d'utilisateurs concernés pour gérer ensemble et supprimer ces contrôles. Regardez « Data.gouv.fr » que l'on gère de façon communautaire à partir des règles du Web Social. Six mois, deux développeurs, pour réécrire le site avec de l'Agile … Aujourd'hui cela coûte 120 k€ (embauche de deux développeurs). Trois fois plus de visiteurs, des techniques Web impensables avant…. En 6 mois, nous avons développé le Marché Public Simplifié avec la même méthode pour répondre aux Appels d'Offres sans renvoyer de documents pour répondre et sans tous les documents demandés (K-bis, adresse fiscale, interdits BdF …). Toutes les technologies que nous avons utilisées l'ont été accompagnées des développeurs géniaux. Peut-être ne peut-on pas transmettre la technologie sans les hommes qui vont avec !

Maintenant, nous rêvons que le décret de création du Chief Data Officer de l'Etat soit promulgué dans la semaine … Comme on n'aime guère les anglicismes, on traduira CDO par Administrateur Général des Données. C'est parti d'un constat de quelques limites de l'Open Data : c'est bien de partager des données mais si l'on ne peut jamais savoir d'où elles viennent et comment elles sont construites, si l'on ne peut pas de temps en temps faire converger deux SI c'est un peu du gâchis. Nous passons notre vie dans les ministères à promouvoir ouverture, qualité, production et convergence de la donnée, complétude, vitesse de rafraichissement…
Un deuxième sujet concerne la gouvernance de ces données : « vous nous avez donné de nouvelles responsabilités en département : les routes, la qualité, mais nous n'avons pas de données y afférentes ». Et là cela devient très compliqué : la donnée est collectée par un gendarme, agrégée en région, elle part en central où elle est requalifiée et

normée, puis validée avant de revenir 18 mois plus tard ! Sans oublier les chamailleries et mesquineries. De plus, une réforme fiscale ne peut se modéliser ! L'an dernier nous avons relevé le défi avec une équipe de professionnels qui ont codé les 400 règles existantes. Il a fallu nettoyer, (par exemple il existe trois définitions de « la famille » avec des assiettes et des taux différents…) Finalement, nous avons réussi à bâtir un modèle complet et maintenant on s'attaque à la simplification de la fiche de Paie.
Pôle Emploi n'a pas encore commencé le Datamining des Entrées/Sorties du chômage pour savoir si les stages servent à quelque chose, si les aides servent à quelque chose, la probabilité que si j'accepte 10% de baisse de salaire je trouverai du travail deux fois plus vite … Toutes ces valeurs sont essentielles et je vais devoir animer cette filière. Sachant qu'il n'est pas souhaitable que les Services du premier ministre assument trop de risques opérationnels car le Premier Ministre joue un rôle d'arbitre et il ne peut donc jouer. Nous stimulons le mouvement par la preuve, en allouant des ressources et en faisant se rencontrer des chercheurs.

L'AGD, Administrateur Général des Données, tarde à être nommé car depuis qu'il a été défini il a croisé deux Premiers Ministres, trois remaniements et un Secrétaire Général en 6 mois ! L'AGD connaîtra les données, se prononcera sur leur qualité au travers d'un rapport annuel, produira les données si nécessaires (cf. le problème de la Base Nationale des Adresses Géolocalisées : l'Etat n'a pas l'adresse complète des bâtiments, il s'adresse aux quatre détenteurs actuels : La Poste, la DGFIP, l'IGN et l'INSEE qui ont tous des bases incomplètes (12 à 14%), des bugs et ne sont pas Open Data. Peine perdue mais personne n'accepte d'ouvrir sa base. Aussi, nous avons construit un accord avec les cartographes de Wikipédia. En six mois ils ont fait 80% du travail, et nous allons voir si nous pourrons finir seul!

Avant de terminer, je dois vous dire qu'aujourd'hui il a quatre familles de questions à affronter :
1/ La gouvernance et la gestion des secrets ce qui n'est pas trivial. Ces secrets sont légitimes et *silotés* : secrets médical, des affaires, fiscal …Trop mélanger ces secrets est hasardeux car nous risquons de perdre la culture métier ; nous n'avons pas le droit de prendre toutes les données mais de savoir ce qu'ils ont dans le système ;
2/ L'articulation avec la DSI : avec la DISIC nous nous entendons très bien mais le Chief Data Officer a un pied du coté des systèmes d'information et personne ne sait vraiment ou est la frontière. La DISIC prépare une stratégie technologique intégrée pour l'Etat qui devient quand même une stratégie de plate-forme, car ils ont entendu nos besoins, nos demandes et nos pratiques, cela contamine la conception du système d'information.
3/ J'ai un peu vendu mon *affaire* mais il va falloir s'y mettre et il faut un profil de personne un peu spéciale. C'est un hacker de la donnée qu'il faut … Faire naitre une vraie équipe qui s'intègre vraiment dans une organisation et produit des résultats qui nourrissent la décision. Ceci est un grand défi.
4/ Enfin il reste la question de la *scalabilité*. Aujourd'hui c'est à taille humaine et il va falloir contaminer des organisations entières et croitre avec plus d'Agile et plus de réactif car il reste toujours 5 millions de fonctionnaires à payer, 40 millions de feuilles d'impôts, du zéro défaut parfois….

Débat

Intervenant : On voit l'intérêt de l'Open Data pour l'état mais y a t il des applications dans le monde privé ?
Henri VERDIER : C'est balbutiant mais on voit des tas de choses qui n'en sont pas si loin. Il y des vraies stratégies de communication fondées sur la transparence. Dans les grandes sociétés qui font de l'Open Innovation, on commence quand même à voir des initiatives assez fortes (A la SNCF, la Poste, la RATP), de partage de données. Ce sont souvent des anciennes entreprises publiques qui ont surement davantage la culture du partage. Elles écoutent et anticipent davantage …

Int : Y a t il des réflexions sur obligation de publication des données ?
HV : Je pense que oui. L'Etat fabrique le bien commun, ce qui fait le lien social, la puissance de notre économie, le *commun knowledge*. Par exemple, nous avons mis en Open Data la base des médicaments mais pas les mises sur le marché (car c'est secret industriel !). De même, les assureurs connaissent le palmarès des voitures les plus dangereuses.

Int : Pensez-vous que l'ingénierie autour de la data va permettre de remplacer certaines décisions ?
HV : Je ne sais pas. On a souvent cru que cela nourrissait les décisions stratégiques, mais en fait les endroits ou il y a eu un impact puissant c'est lorsque ça aide dans des micros décisions.

Int : Vous avez évoqué les canulars comme étant un élément perturbateur de la mécanique. N'est ce pas un élément dangereux dans l'Open Data dans le futur ?
HV : Pour répondre sur data.gouv.fr, il y a des données sur lesquelles on s'engage. Pour d'autres données on dit que c'est sous la responsabilité de X ou Y. Ces systèmes sont résilients et on arrive à s'en sortir si on a conçu des règles avec des alertes et des règles de vérification ; On doit tout surveiller tout le temps.

Int : Vous avez parlé d'Open Data française, mais y a t il des enjeux pour normer les données ?
HV : Il y a 3 ou 4 sujets à ne pas mélanger : La première est l'indexation. Certaines administrations devaient indexer une première fois pour Eurostat, puis pour les archives de France et, enfin, pour nous. Ce n'est pas possible ! Après il y a la standardisation et le Web sémantique. Je préfère dire aux gens : commencez ! Partagez en format brut et si quelqu'un veut faire converger les standards, il le fera, car sinon nous perdrons vraiment trop de temps.
Autant que possible, je cherche les standards de fait, les standards minimums et je regarde les indexations le plus tard possible, et j'ai peut être tort !

Int : Le CDO est-il le futur du CIO, et quelle est sa valeur ajoutée dans une organisation ?
HV : Je ne sais, je me méfie des appellations. Ce qui compte c'est que les fonctionnalités soient remplies harmonieusement. Le bon équilibre sera l'articulation harmonieuse entre quelqu'un très côté usage, design, interface avec le grand public, et un autre très côté back office, rationalisation, agilité. A la fin un pôle rigueur et un pôle agile et pourquoi pas le tout dans une DSI ?

Int : Quel est le plan de communication pour faire savoir au grand public que ça existe ?
HV : Aux US, Obama peut en parler une heure sans papier aux journalistes devant la Silicon Valley, qui marche mieux qu'Hollywood et Wall Street … mais chez nous le politique ne comprend pas tout !…et cela n'intéresse pas les journalistes.

Présentation de l'orateur

Henri VERDIER

Après avoir été le directeur général de la société Odile Jacob Multimédia, où il a notamment développé un ensemble de supports pédagogiques pour la « Main à la Pâte », Henri VERDIER rejoint en 2007, Lagardère Active comme directeur chargé de l'innovation. En 2009, il rejoint l'Institut Télécom comme directeur de la prospective, chargé de la création du Think Tank « Futur numérique »
Il est co-fondateur de la société MFG-Labs, qu'il quitte en 2012 et qui est acquise par Havas Media en 2013.
Membre fondateur du « Pôle de compétitivité Cap Digital », il en exerça la vice-présidence de 2006 à 2008, avant d'en être élu président du Conseil d'administration de 2008 à janvier 2013.
Il dirige, depuis janvier 2013, Etalab, le service du premier ministre chargé de l'ouverture des données publiques. Sous sa direction, Etalab a développé une nouvelle version du portail d'open data français « data gouv.fr », qui héberge de nombreuses données publiques. Cette version autorisant les citoyens à enrichir les données publiques où à partager leurs propres données a été qualifiée par le blog TechPresident de « Première mondiale »