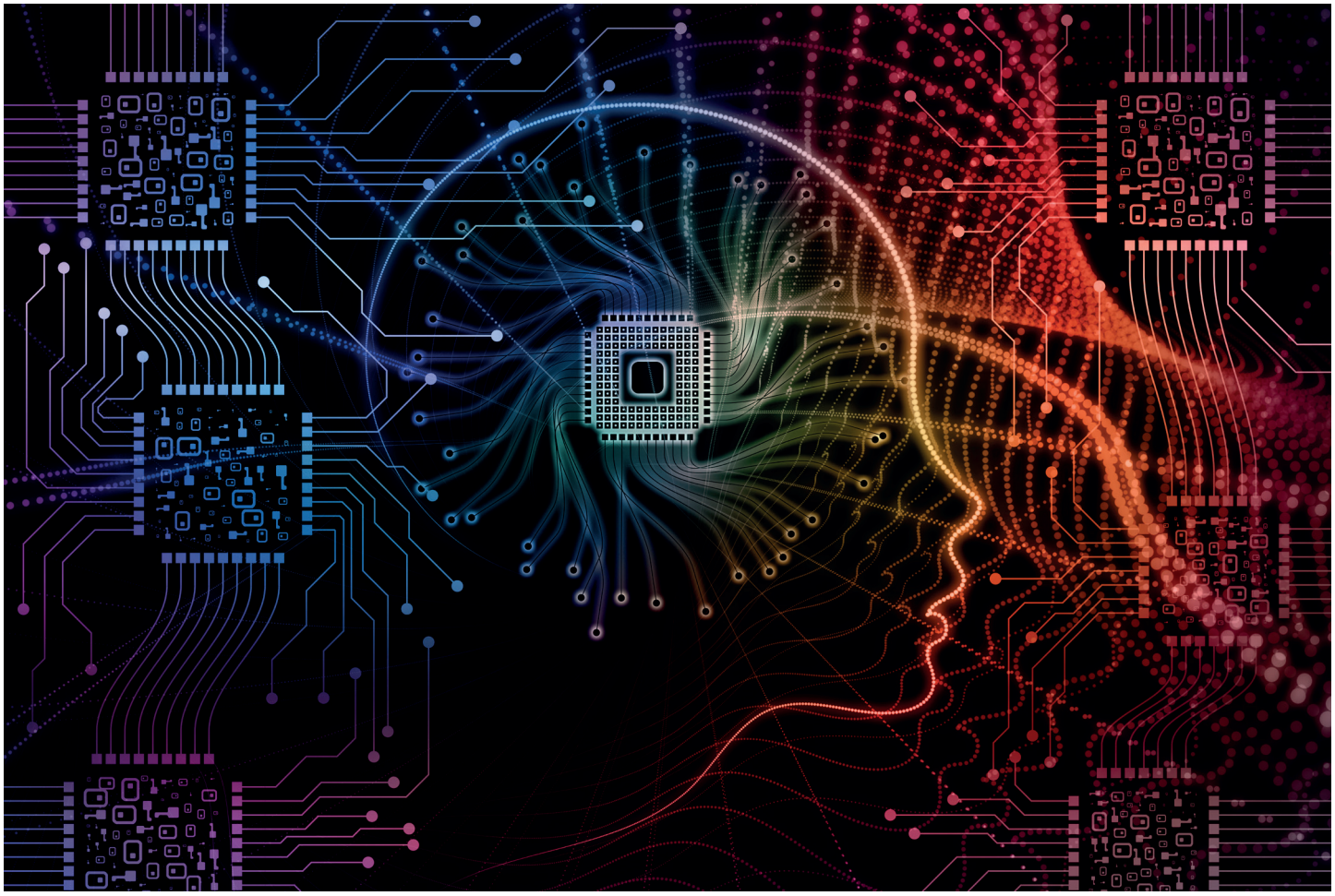
À QUOI SERVENT LES DATALABS ?

SYNTHÈSE DU PETIT DÉJEUNER DÉCIDEURS-CHERCHEURS DU 23 MARS 2021



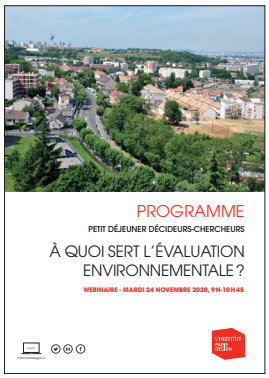
JUILLET 2021
30.20.01

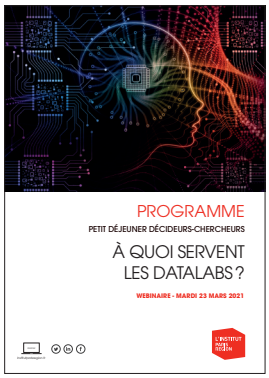
www
institutparisregion.fr

L'INSTITUT PARIS REGION

Cette synthèse présente les principaux enseignements du petit déjeuner « À quoi servent les datalabs ? »

Elle s'inscrit dans le thème transversal des petits déjeuners décideurs-chercheurs 2020-2021 sur les territoires en transition :
- Comment ça marche en Île-de-France ? (17-06-2020)
- Cohabiter avec les animaux sauvages en milieu urbain (30-09-2020)
- À quoi sert l'évaluation environnementale ? (24-11-2020)
- À quoi servent les datalabs ? (23-03-2021)





Retrouvez les ressources documentaires (podcast, diaporamas, bibliographie, etc.) sur le site de L'Institut Paris Region:
https://www.institutparisregion.fr/petits-dejeuners-decideurschercheurs.html

Directeur général : Fouad AWADA
Synthèse rédigée par Antoine COURMONT, Brigitte GUIGOU, Guillaume LECOEUR, Dany NGUYEN-LUONG à partir d'une retranscription de Béatrice MERCIER.
Coordination : Brigitte GUIGOU
n° d'ordonnancement : 30.20.01

Crédit photo de couverture : Agsandrew/shutterstock.com

À QUOI SERVENT LES DATALABS ?

La disponibilité d'une masse de données individuelles, à des échelles spatiales fines, ouvre de nouveaux terrains et sujets d'études aux acteurs de la ville. Leur analyse, via de nouvelles méthodologies quantitatives, contribue à enrichir les politiques publiques et nourrir la décision, notamment en matière de mobilité. Pourtant l'organisation et l'usage de cette masse de données posent nombre de questions. Quel cadre juridique, économique, éthique et démocratique construire ? Comment les acteurs publics peuvent-ils avoir accès à ces données, aux mains d'opérateurs privés ou publics ? Comment trier, traiter, agréger, représenter et donner du sens à ce torrent d'information dans le cadre d'un datalab ? Comment coupler ces données avec celles issues d'enquêtes quantitatives classiques ? Quel mode d'organisation et compétences privilégier, notamment dans les agences d'urbanisme ?

Pour répondre à ces questions au cœur des pratiques et des productions de L'Institut Paris Region, un chercheur et un décideur croiseront leurs points de vue.

PROGRAMME
DU 23 MARS 2021 (EN WEBINAIRE)

9 h 00 : OUVERTURE
Fouad AWADA, *directeur général de L'Institut Paris Region*

- **Introduction :**
  Dany NGUYEN-LUONG, *directeur du département Mobilité Transports à L'Institut Paris Region*

- **Organisation et animation :**
  Brigitte GUIGOU, *chargée de mission partenariat recherche à L'Institut Paris Region*

9 h 30 – 10 h 30 : INTERVENTIONS ET QUESTIONS/RÉPONSES
- **Antoine COURMONT,** *chercheur en science politique, directeur scientifique de la chaire Villes et numérique de l'école urbaine de Sciences Po*
- **Guillaume LECOEUR,** *responsable du pôle Données et Innovation, SNCF Réseau*

À QUOI SERVENT LES DATALABS ?

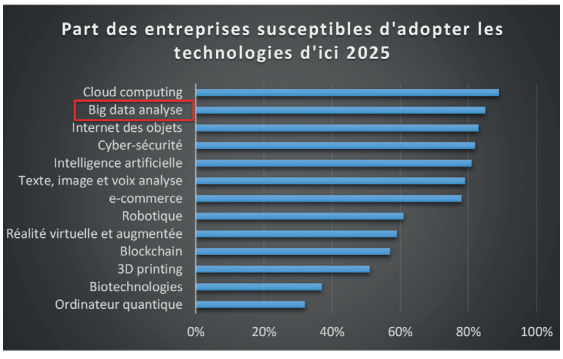
Ouverture

Fouad AWADA,
directeur général de L'Institut Paris Region
Tout le monde connait l'importance de l'information et de la donnée dans le fonctionnement des entreprises. Nous avons besoin de savoir pour agir. Combien de clients pouvons-nous toucher ? À quelle distance ? Quels sont leurs revenus ? Depuis longtemps les entreprises mobilisent des données pour répondre à ces questions. Mais depuis peu on s'intéresse à la valeur de ces informations pour soi, mais aussi pour les autres. Certains d'entre vous, parmi les plus âgés, se souviennent peut-être des premières cartes utilisées par les randonneurs. Il s'agissait de cartes d'État-major de l'armée, détournées vers d'autres usages. Dans nos métiers, il nous arrivait régulièrement d'utiliser l'annuaire téléphonique pour réaliser des enquêtes, tirer des échantillons. Si ce détournement n'est pas nouveau, nous sommes entrés aujourd'hui dans une toute autre dimension en raison de la croissance exponentielle des données produites grâce au numérique. Certains en ont fait un business, c'est le cas de l'économie biface qu'on appelle aujourd'hui les plateformes qui collectent les données et les vendent. Les sociétés comme Orange, Coyote, SFR, Carrefour, etc., disposent de données qui, pour certains, ont un autre usage et qui donc vallent de l'or. Ces sociétés se sont questionnées sur la manière de les réemployer. Il s'agit donc de considérer la donnée, celle qu'on produit ou celle qu'on collecte, comme un actif à valoriser et d'avoir une réflexion stratégique sur la manière d'y parvenir. À l'heure du Big data et des capteurs, peut-on se passer d'une telle démarche dans n'importe quelle entreprise ? L'idée de créer un Datalab à L'Institut Paris Region s'inscrit dans ce contexte.
L'objectif de ce petit déjeuner est de nous éclairer sur cette thématique et je remercie tous les intervenants présents ce matin.

Brigitte GUIGOU,
Chargée de mission partenariat recherche, L'Institut Paris Region
En effet l'organisation et l'usage de cette masse de données individuelles, à des échelles spatiales fines, contribuent à enrichir les politiques publiques et nourrir la décision. Mais ils posent aussi nombre de question aux acteurs de la ville, Quel cadre juridique, économique, éthique et démocratique construire ? Comment les acteurs publics peuvent-ils avoir accès à ces données, aux mains d'opérateurs privés ou publics ? Comment trier, traiter, agréger, représenter et donner du sens à ce torrent d'information dans le cadre d'un datalab ? Comment coupler ces données avec celles issues d'enquêtes quantitatives classiques ? Quel mode d'organisation et compétences privilégier, notamment dans les agences d'urbanisme ?

Pour répondre à ces questions au cœur des pratiques et des productions de L'Institut Paris Region, un chercheur en science politique, Antoine Courmont, et un décideur, Guillaume Lecoeur responsable du pôle Données et Innovation, SNCF Réseau, croiseront leurs points de vue. Au préalable, Dany Nguyen-Luong, directeur du département Mobilité Transports à L'Institut Paris Region, proposera une intervention de cadrage.

Dany NGUYEN-LUONG,
Directeur du département Mobilité Transports, L'Institut Paris Region
Je vous propose un cadrage sur le concept de Datalab que j'illustrerai par un cas d'usage, le tableau de bord de la mobilité mis en place par L'Institut en novembre 2020.
Un Datalab se base sur l'utilisation du Big data mais aussi sur celle d'autres technologies. Le graphique joint montre la part des entreprises susceptibles d'adopter d'ici 2025 des technologies telles que « le Cloud computing », « le Big data », « l'Internet des objets », etc., jusqu'à l'Ordinateur quantique. Le Big data occupe une bonne place, puisque 85 % des entreprises ont l'intention d'exploiter le Big data d'ici 2025. Le Big data est très lié aux autres technologies, au Cloud, à l'Intelligence artificielle, à l'Ordinateur quantique, au Calcul haute performance, à la Modélisation simulation... Ces technologies sont indissociables.



Source : The Future of jobs report - word economic forum 2020)

Le Big data est, depuis une vingtaine d'années, caractérisé par les fameux 3 V (volume-variété-vélocité). Dans le domaine de la mobilité, on cherche des données d'usage et de fréquentation, beaucoup plus difficiles à collecter que des données d'offres.

Je retiendrais 4 caractéristiques du Big data.

- **Le volume des données.** On parle aujourd'hui de gigaoctet, bientôt de teraoctet, de petaoctet (soit 10 puissance 15), de yottaoctet (soit 10 puissance 24) et puis un jour lointain, on parlera de google (soit 10 puissance 100).
- **La variété des données.** Ce sont les traces numériques des déplacements géolocalisés avec des données, de type GPS, d'opérateurs téléphoniques « les FMD » ou de télébillétiques.
- **La rapidité d'accès en « quasi temps réel ».** On trouve des données en Open data disponibles à J+1 ; par exemple, pour les données du trafic routier à Paris. L'idéal ce sont ces plateformes d'accès aux données à un format Opendatasoft avec possibilité de filtrage. Malheureusement, ce n'est pas toujours le cas. L'Opendata est essentiel à la réussite des datalabs. Nous avons fait beaucoup de chemin depuis l'initiative Etalab il y a 10 ans, d'abord avec la loi pour le numérique, puis avec le récent rapport Bothorel qui recommande un meilleur partage des données entre acteurs publics et la nécessité qu'ils puissent accéder à des données produites par le privé lorsque celles-ci sont considérées d'intérêt général. La loi LOM va aussi dans le sens d'une ouverture de données. La crise Covid a également montré le besoin d'ouvrir les données pour permettre aux citoyens, à la société civile, aux chercheurs ou aux médias de s'approprier ces données pour aider les pouvoirs publics à mieux suivre l'évolution de la pandémie.
- **Données horodatées et historisées.** Elles permettent de faire des analyses d'évolution temporelle qui font l'objet de graphiques dans notre tableau de bord de la mobilité en ligne.

Pour mettre en place un Datalab, il faut avoir à l'esprit le triptyque « collecter, traiter et partager ». Dans le domaine de la mobilité, il y a les données classiques provenant des enquêtes traditionnelles de type ménage-déplacement, migrations alternantes du recensement, mais aussi des enquêtes origines destinations dans les gares. Ces enquêtes classiques sont indispensables. Ce sont des références, la matière première pour les analystes de la mobilité et les modélisateurs. Mais elles souffrent d'une trop faible fréquence et d'un coût élevé. En période de crise, dans un moment où l'on cherche à suivre la mobilité en « temps réel », elles sont rapidement dépassées. D'où l'intérêt de données de type Big data, ces données innovantes comme celles de Traces numériques, de télébilletique et espérons un jour, les données issues des capteurs 3D dans les gares et dans les matériels roulants. L'avantage des Big data, ce sont les 3 V, leur fraîcheur, la connaissance fine des origines destinations, et elles sont gratuites et disponibles en Opendata. Elles ont également des limites dont il faut avoir conscience, par exemple l'absence d'information sur les profils sociaux-économiques des usagers, ou l'absence de connaissance précise sur les motifs des déplacements. Se pose aussi la question du redressement des données lorsque la base n'est pas exhaustive.
Le graal d'un datalab en mobilité est une base de données au croisement entre les données classiques et le Big data. Cette base existe, tout le monde l'alimente sans forcément le savoir et il faut juste savoir comment la récupérer. C'est la Google Maps Timeline.

On a parlé des données temps réel. L'INSEE parle de données haute fréquence mais il y a aussi des données alternatives qui sont utilisées à d'autres fins que celles pour lesquelles elles ont été produites. Par exemple, les données des opérateurs téléphoniques utilisées pour suivre la mobilité par « grande masse » ou par « origine destination ». Un autre exemple récent est l'analyse des eaux usées, pour évaluer l'évolution de la pandémie du Covid avec plusieurs jours d'avance par rapport au test. Dans notre tableau de bord « Mobilité », nous avons essayé d'utiliser ces données alternatives pour suivre les touristes internationaux à partir des données de transaction de cartes bancaires, en remontant les terminaux de paiement. Ce sont des données en Open data de la BPCE. Il y a d'autres exemples comme le suivi de l'activité économique à partir des données de recettes TVA, celles sur la consommation électrique dans le bâti, les images satellitaires pour identifier le taux d'occupation dans les parkings des centres commerciaux. On peut dire que l'ortho-photo c'est une photo, alors que les images satellitaires sont un film qui permet de suivre l'évolution. Parmi les autres données alternatives il y a base de données DVF sur les transactions immobilières en open data avec des mises à jour tous les six mois. Elle permet de suivre les effets urbains des transports ou les effets de relocalisation résidentielle. La mise à jour des données de DVF du second semestre 2020 est attendue en avril. Pour suivre les déplacements pour motif santé-accompagnement, la méthode classique serait de lancer une enquête auprès de la population et des professionnels de santé.

Aujourd'hui, on essaie de recueillir les données auprès de l'Assurance Maladie sur la téléconsultation. L'intérêt d'un datalab – qui permet de recouper toutes ces données au sein d'une même structure – est de créer un nouveau réflexe d'utilisation de données alternatives. Ces deux méthodes, classique et datalab, sont complémentaires.
Dans le tryptique « collecter – traiter – partager », la phase de traitement est essentielle. Il faut penser « traitement, nettoyage, redressement ». Il y a une panoplie de traitements mathématiques et statistiques plus ou moins complexes. Cela va des tableaux croisés dynamiques aux méthodes de régression linéaire ou logistique et aux méthodes de classification (ACP, hiérarchique). Depuis quelques années, pour faire du prédictif, on utilise des méthodes de machine learning, notamment avec l'algorithme « Random forest », et des méthodes de Deep learning par réseau neurone artificiel avec l'algorithme de « rétropropagation du gradient ».

La phase de partage de données est aussi importante. « Partager, c'est valoriser les données ». La visualisation des données fait partie de la science des données. Le choix de la représentation graphique est important, comme le montre le site « Covidtracker » qui met bien en valeur les données en open data d'épidémies. Si on parle d'or noir pour la donnée, c'est parce que les outils de data visualisation sont aux données ce que le raffinage est au pétrole. Il existe des dizaines d'outils de visualisation, par exemple Infogram, Tableau, Qlik View, Spotfire, Saagie, Power BI, etc. À L'Institut, nous utilisons Infogram. Certains outils parviennent à automatiser la chaîne collecter-traiter-partager. Par exemple, pour répondre à une requête en langage naturel du type « quelles sont les ventes de voitures en Espagne au dernier trimestre 2020 ? », le logiciel va chercher la base de données en open data, la traiter et afficher le graphique de résultats en optant pour la meilleure représentation. C'est ce qu'on appelle une « data virtualisation ».

Qu'est-ce qu'un datalab ?

Au sein d'une entreprise, un datalab est une structure dédiée à la collecte, l'exploitation et la visualisation des données. Tout ce qui tourne autour de l'innovation et de la donnée permet de croiser les expertises des experts métiers et des data scientists. Un datalab est une structure transversale apportant de la souplesse et de la réactivité aux organisations sans en remanier la hiérarchie ou l'organigramme. C'est aussi l'opportunité d'intégrer de nouvelles compétences telles que les data scientists. Un datalab peut fonctionner comme une startup au sein de l'organisation. Il permet de faire travailler ensemble les services informatique et communication. C'est une démarche « agile », où l'on choisit un sujet et où l'on procède par essais-erreurs. C'est ce que nous avons fait avec notre tableau de bord de la mobilité. L'étape suivante sera d'automatiser la chaîne « collecter-traiter-visualiser » en utilisant des API permettant de télécharger des jeux de données à la volée et de les traiter. Un datalab offre des perspectives d'innovation et de transformation dans les agences d'urbanisme, les services techniques de l'État et des collectivités. Cela bouscule évidemment les habitudes de travail.

Pour conclure, je dirai un mot sur l'exemple du projet NEON. C'est un projet fédéral américain dans le domaine de l'écologie : National Ecological Observatory Network (NEON). Les données sont collectées par des dizaines de milliers de capteurs répartis sur tout le territoire dans les domaines de la biodiversité, du changement climatique, de l'hydrologie, des maladies infectieuses, de l'artificialisation, de la faune et la flore, de la météo, etc. Elles sont mises en ligne et disponibles gratuitement. Le financement est sanctuarisé sur trente ans depuis 2018 avec 80 millions de dollars par an. Je vous invite à consulter leur site internet (https://www.neonscience.org/). On peut rêver d'un observatoire de ce type-là en Île-de-France dans le domaine des transports, de l'urbanisme et de l'environnement. Notre tableau de bord de la mobilité est la première pierre de ce vaste édifice.

Brigitte GUIGOU
Qu'est-ce que la base de données Maps Timeline ?

Dany NGUYEN-LUONG
Maps Timeline est la base de données de Google qui fonctionne lorsque l'on active sans le savoir la fonction de géolocalisation. Tous nos trajets sont tracés, enregistrés par Google. Maps Timelines garde l'historique des trajets dans le détail sur le modèle d'une enquête ménage déplacement avec pour chaque déplacement l'heure de départ et d'arrivée. Sur cette base il est possible de faire des hypothèses au motif à la destination et d'arriver à détecter le mode de transport utilisé. Cette base de données des déplacements désagrégés existe, c'est pourquoi L'Institut souhaiterait en récupérer un échantillon.

Antoine COURMONT,
Chercheur en science politique, directeur scientifique de la chaire Villes et numérique, École urbaine de Sciences Po

Caractère politique des données et recompositions des pouvoirs associés à leur production et à leur utilisation

La donnée sous forme d'information statistique, cartographique ou d'état civil a joué un rôle central dans le processus d'étatisation des sociétés. C'est-à-dire dans la capacité de l'État moderne à imposer une représentation à l'ensemble de la société, à mettre en œuvre des politiques publiques. La construction de l'État-Nation est indissociable de la capacité de l'autorité politique à acquérir une forme de monopole sur la production de données légitimes et à influer sur notre capacité à voir le monde au travers de catégories étatiques. C'est ce que le sociologue Luc Boltanski appelle le pouvoir sémantique des institutions, leur capacité à représenter des phénomènes sociaux et surtout à coordonner les acteurs à partir de cette définition de la réalité.
Les processus sont inégaux selon les pays qui ont une capacité plus ou moins importante à produire ces informations et à les imposer. À l'inverse, ne pas produire de données pour un acteur politique public peut être un excellent moyen de produire de la méconnaissance et de ne pas gouverner certains secteurs d'action publique. Notamment, des travaux de recherche mettent en évidence le fait que les pouvoirs publics produisent volontairement de la méconnaissance sur certains phénomènes sociaux. Je pense notamment aux travaux de Thomas Aguilera sur les habitats informels souvent illégitimés par les autorités publiques en France et à l'étranger. La connaissance de ces phénomènes est faible, ce qui empêche leur mise à l'agenda politique et leur prise en charge.

En suivant les capacités de différents acteurs à produire des données, on peut observer les recompositions dans la gouvernance des territoires. Bien avant le numérique, à partir des années quatre-vingt en France avec les lois de décentralisation, il y a eu une montée en puissance des collectivités locales, notamment des structures intercommunales. Elle s'est accompagnée d'une capacité à produire des données sur son territoire, à le représenter notamment au travers de la mise en place de système d'information géographique ou en association avec des tiers telles que les agences d'urbanisme. Là où les villes étaient dépendantes des services de l'État et de leur capacité à produire des informations sur leur territoire, elles ont gagné en autonomie. Elles ont été en capacité partielle de représenter leur territoire et d'agir, même si l'État conserve de nombreuses prérogatives, comme certaines entreprises incontournables dans la gouvernance urbaine.
Les collectivités sont inégalement dotées en termes de capacité de production de données. Cela dépend aussi des secteurs d'action publique. Par exemple, les collectivités se sont longtemps désinvesties du secteur énergétique qu'elles ont laissé aux mains des énergéticiens. Tout récemment, des collectivités souhaitant mettre en place des politiques publiques sur leur territoire en matière de stratégie énergétique ont cherché à récupérer et utiliser ces données.

L'attachement des données et les difficultés qui peuvent émerger quand il s'agit de les partager, les mettre en circulation et les utiliser à des fins alternatives

La recomposition des relations de pouvoir entre échelles gouvernementales peut aussi s'analyser par la production de données. L'entrée de nos sociétés dans l'ère du numérique et du big data vers la fin des années 2000, est caractérisée par une capacité accrue d'un ensemble d'organisations publiques, privées ou de la société civile, pour produire, stocker, traiter et faire circuler la donnée. Là où auparavant les coûts et les investissements nécessaires à la production de données étaient réservés à de grandes administrations, aujourd'hui les start-ups, les citoyens, les organisations peuvent produire de la donnée à des coûts beaucoup plus faibles.
Dès lors, tous ces acteurs ont été en mesure de produire des informations fournissant des représentations alternatives de nos sociétés et de nos territoires. On peut prendre l'exemple d'Open TripMap, un projet de cartographie libre, le Wikipédia de la cartographie. Cela aurait été inimaginable avant l'essor du numérique, qui a impliqué la capacité de tout à chacun à acquérir un GPS. Ce projet a permis de cartographier de nombreuses zones, par exemple des bidonvillesou des quartiers informels qui ont acquis une existence, une forme de légitimité et donc une mobilisation politique possible à partir de cette production de données alternatives.

Du côté de la société civile et des citoyens

Les citoyens sont aujourd'hui en mesure de produire des statistiques qu'ils peuvent opposer à l'État. On l'observe avec le mouvement des capteurs citoyens. Ce sont des citoyens qui par exemple, décident de mesurer la qualité de l'air et de produire des mesures alternatives pouvant s'opposer aux mesures officielles. Ces mesures alternatives sont de nouvelles formes d'engagement politique.

Du côté des acteurs privés

Cette capacité accrue en matière de production de données est aussi investie par les acteurs privés, ce qui peut conduire à des conflits de régulation assez forts. C'est le cas par exemple avec l'entreprise Waze, qui propose une application de calculateur d'itinéraires. Les données proviennent des usagers, ce qui rend l'application indépendante des pouvoirs publics. Ses modalités de calculs algorithmiques conduisent à des reports de trafics dans des quartiers résidentiels ou des zones peu fréquentées par les automobilistes. Cela provoque des conflits avec les autorités en charge de la régulation de la circulation automobile. Les pouvoirs publics se voient dépourvus de leur capacité à gouverner et perdent en quelque sorte la maitrise de la représentation de leur territoire.

Un des grands enjeux aujourd'hui en matière de transformation de gouvernance est la mise à l'épreuve de ce pouvoir sémantique des institutions publiques. L'État, les collectivités ont perdu leur monopole de production de données à partir desquels des individus vont se coordonner, que ce soit à l'échelle nationale avec l'identité numérique, en partie aux mains de Facebook ou de Google versus l'état civil précédemment, ou à l'échelle locale avec les listes de meublés touristiques que possédent Airbnb ou les listes de véhicules avec chauffeur possédées par Uber. Il faut signaler l'importance pour les pouvoirs publics de regagner une expertise en matière de production et de traitement de données pour conserver la maitrise de la représentation des territoires et des phénomènes sociaux et in fine pour conserver la maitrise des politiques publiques sur leur territoire. C'est à ce titre que les datalabs peuvent jouer un rôle assez crucial.

La mise en œuvre des datalabs

L'objectif des datalabs repose souvent sur un dispositif, une infrastructure technique et une forme de plateforme des données. Le datalab permet d'agréger une architecture plus ou moins centralisée selon des procédures plus ou moins automatisées et des formats devant être standardisés et exploitables de données provenant de différentes sources et organisations. Auparavant ces données quittaient rarement leur système d'information d'origine.

Mettre en œuvre cette circulation et ce partage de données en amont de tout traitement demande des investissements et un travail non négligeable. On peut avoir l'impression que c'est facile, qu'il suffit de faire un copié/collé dans une base d'information métier du producteur vers un système d'information de diffusion. Dans les faits, c'est bien plus compliqué car les données sont loin d'être des éléments neutres et immatériels qui circulent aisément. Elles sont, au contraire, solidement attachées à de vastes infrastructures sociotechniques composées d'organisation, de modèles économiques, de cadre juridique, de systèmes d'information, de format voire même de culture métier. Pour les mettre en circulation, il est nécessaire de défaire ces liens. Cela requiert un travail considérable notamment parce que les défis techniques et organisationnels sont intimement liés. Il faut insister sur l'aspect organisationnel qui est tout aussi important que l'aspect technique.

En matière technique, il y a les activités de standardisation, de partage, d'agrégation de bases de données métiers. Cela génère des problématiques matérielles de mise en cohérence de format. L'enjeu est aussi de réussir à partager et extraire des données de systèmes d'information métier qui peuvent être des systèmes propriétaires, anciens, qui ne sont pas conçus pour cela et qui nécessitent pour être mis en circulation des investissements parfois importants. Il est aussi nécessaire d'enrôler les organisations productrices de données pour qu'elles acceptent de les mettre à disposition, ce qui ne se fait pas sans réticence.

Un autre enjeu est l'attachement de données économiques avec des modèles d'affaires pouvant être associées à des données qui contraignent leur mise en circulation et leur partage. Ces données sont également attachées à des cadres juridiques et réglementaires. Quand elles rentrent dans le cadre du RGPD (Règlement Général sur la Protection des Données), un travail conséquent d'anonymisation est nécessaire avant tout partage. C'est le cas des données bancaires ou de transport de Google.

Les données sont porteuses d'un héritage qui contraint leur utilisation. Elles ont été produites pour une finalité et véhiculent une représentation de l'espace qui rend plus difficile leur usage à des fins alternatives. Par exemple, il y avait un jeu de données sur les toilettes publiques mis à disposition sur la plateforme open data du Grand Lyon. Ce fichier, produit par la direction de la propreté de la Métropole de Lyon, recensait et cartographiait ces installations. Quelques semaines après sa diffusion, l'équipe en charge de l'open data a reçu un email d'un utilisateur mécontent parce que seules trois toilettes publiques était listée sur le territoire de la ville de Lyon. D'après la direction de la propreté ce problème était dû au mode de recensement qui, dans la commune de Lyon, identifiait uniquement les toilettes nettoyées par des agents. Hors un grand nombre de toilettes publiques sont des sanisettes autonettoyantes. De plus la direction de la propreté de la Métropole ne disposait pas d'information sur leur localisation, l'information étant

détenue par la ville de Lyon qui prend les arrêtés d'occupation de l'espace public pour leur installation. Cet exemple montre que n'importe quelle donnée véhicule une définition de ce que l'on souhaite représenter en fonction de l'usage que l'on souhaite en faire. L'utilisateur peut faire ce travail d'enrichissement de données mais il faut qu'il aille rechercher les données auprès de différents producteurs. Cela met en évidence ce travail – important mais souvent négligé – de sourcing, de nettoyage, d'enrichissement de la donnée, nécessaire pour l'utiliser à des fins alternatives. Cela explique les limites des portails open data d'aujourd'hui.

Construire une expertise des compétences et nouveaux métiers

Des nouveaux métiers liés au traitement de la donnée émergent dans les organisations privées et publiques :

- les chief data officer, administrateur général de données, les chefs de projets data, etc. ;
- les postes sur la transformation des systèmes d'information : les data architectes, les data ingénieurs ;
- des nouveaux métiers autour des outils et des pratiques et des métiers d'analyses de données : les data analystes, les data scientistes.

Ces profils restent rares et sont donc très recherchés. Leur bonne intégration dans les structures territoriales nécessite de construire une complémentarité avec les expertises métier traditionnelles. Assez techniques, ces profils ont peu de connaissances des enjeux urbains, ce qui peut provoquer des difficultés de compréhension et de construction d'un langage commun avec les experts métier.

J'ai par exemple mené une enquête pour la Ville de Paris sur le réaménagement de la place de la Nation. La ville avait décidé d'installer, en partenariat avec un ensemble de start-up et avec l'entreprise Cisco, des capteurs pour comprendre l'usage des modes doux sur la place. Or ces données ont été très peu utilisées par la Ville de Paris en raison des difficultés de compréhension entre les différentes cultures métier avec d'un côté, une culture urbaine d'aménageurs traditionnels, et de l'autre des cultures issues du secteur informatique. Ces obstacles s'aplaniront sans doute avec le temps. En effet l'émergence et l'usage des Big data dans les politiques urbaines sont récents, construire de l'expertise autour de ces nouvelles sources de données prend du temps. Si la question des flux automobiles est bien maitrisée aujourd'hui c'est parce qu'elle est étudiée depuis un siècle, qu'il y a des chercheurs et experts spécialisés. Or jusqu'à présent les expertises sur les modes doux sont rares et les calculs de flux encore plus. Cela prendra du temps pour collecter les bonnes données et savoir les intégrer dans des modèles pertinents pour ces usages.

Brigitte GUIGOU

A-t-on une idée du nombre de collectivités locales qui sont aujourd'hui engagées dans ces démarches de datalabs ?

D'autre part, quels sont pour vous les principaux arguments en faveur du développement des datalabs dans les agences d'urbanisme de développer des datalabs ?

Antoine COURMONT

Pour les collectivités locales, je n'ai pas de recensement exhaustif. Les collectivités sont entrées d'abord par l'aspect Open data et par sa mise sur l'agenda. L'enjeu pour elles est de rassembler des données de leur propre système d'information et de ceux de leurs partenaires publics ou privés, autour de la notion de données d'intérêt général. Le Rapport Bothorel a rappelé qu'on dépasse le caractère public ou privé des données si elles sont d'intérêt général. Cela peut ouvrir sur des questions très conflictuelles en matière d'intérêt général et de données pouvant y être attachées. Aujourd'hui, les métropoles françaises sont engagées dans la mise en place de ce type de plateformes. Ces investissements sont longs et assez coûteux en matière technique et organisationnelle. On n'en est aux prémices et il n'y a pas grand-chose de fait encore.

C'est pareil pour les agences d'urbanisme. Il y a une prise de conscience du risque qu'un certain nombre de données expertise, entre les mains des agences d'urbanisme, se voient concurrencées ou par des acteurs privés. Comment les agences d'urbanisme peuvent-elles se positionner pour ne pas perdre la main et conserver leur maitrise en matière de production d'analyse de données ? C'est tout à fait logique d'investir sur ces nouvelles sources de données, d'essayer de les rassembler et de commencer à jouer avec pour voir ce qu'on peut en tirer. C'est un investissement à long terme, nécessaire pour maintenir une indépendance publique en matière d'expertise des territoires.

Brigitte GUIGOU

Vous avez évoqué la question de la collecte des données. Comment inciter les acteurs privés à transmettre leurs données dans le bon format ?

Antoine COURMONT

Le rapport de force est plutôt défavorable au domaine public, en particulier aux institutions

publiques locales dans la mesure où les grosses plateformes numériques n'ont pas forcément intérêt à récupérer des données du domaine public, sauf dans le domaine du transport. Néanmoins un certain nombre d'obligations sont peu à peu transcrites dans la loi et dans le cadre réglementaire obligeant les acteurs de l'économie numérique à transmettre leurs données aux collectivités. C'est le cas par exemple des locations de meublés touristiques, dont Airbnb qui depuis peu transmet des données chaque année à la Ville de Paris. Cela permet à la Ville de mettre en œuvre la régulation et de repérer les loueurs qui ne respectent pas le seuil des 120 jours. La loi s'accompagne de décrets d'applications sur le format de données. Mais concrétement les données sont fournies dans des formats et des structures différents et il y a toujours un enjeu de standardisation pour ces données des acteurs privés.

Brigitte GUIGOU
Vous avez aussi souligné qu'un certain nombre de données s'appuient sur cette dimension collaborative. Cela pose-t-il des problèmes particuliers ?

Antoine COURMONT
Il est important que les acteurs publics investissent sur l'aspect collaboratif. Cela leur permet d'avoir une représentation dépassant leurs frontières territoriales. Mais les différences de formats d'une ville ou d'une intercommunalité à l'autre ne facilitent pas le rôle des utilisateurs externes. S'appuyer sur des initiatives comme Open Street Map permet un certain nombre de standardisations de données. Cela permet aussi de s'appuyer sur la foule pour produire ces données. Comment peut-on alors s'assurer de la fiabilité des données qui peuvent être modifiées rapidement ? Et à quel point peut-on s'appuyer sur ces données pour mener des politiques publiques ? Il se trouve qu'Open Street Map, en continuité de ce que fait Wikipédia, a mis en place des dispositifs pour identifier certaines modifications malveillantes. La puissance du collectif fait qu'on arrive à avoir des données d'assez bonne qualité et assez fiables. Certaines collectivités commencent à utiliser ces données collaboratives pour représenter des phénomènes assez coûteux à représenter car ils demandent une présence massive sur le terrain.

Brigitte GUIGOU
Le secteur associatif est parfois très actif dans la production de données alternatives, je pense par exemple au baromètre des villes marchables à l'initiative de plusieurs associations piétonnes. Quel rôle jouent, selon vous, ces associations ?

Antoine COURMONT
En effet, j'ai aussi cité l'exemple des capteurs citoyens pour les pollutions atmosphériques ou sonores. Cette capacité accrue à produire des données concerne aussi la société civile. Des groupes militants ou des associations produisent leurs propres représentations et indicateurs. Ils proposent des chiffres alternatifs à ceux des pouvoirs publics pour agréger un certain nombre de collectifs, se mobiliser et faire changer les politiques publiques. C'est assez ancien. Certains sociologues ont appelé cela le « statactivisme », c'est-à-dire, lutter avec des statistiques avec un slogan « un autre nombre est possible ». À l'époque, c'était sur les indicateurs alternatifs au PIB représentant la richesse d'un pays. On a la même chose à l'échelle locale aujourd'hui.

Guillaume LECOEUR,
Responsable du pôle données et innovation, DGEX Solutions, SNCF réseau
Le pôle d'innovation de la SNCF, créé depuis 18 mois, s'inscrit dans des transformations qui se sont succédées depuis des dizaines d'années à la SNCF.

Le pôle innovation et ses missions

Aux origines de la création du pôle, nous avons fait le constat que la transformation numérique engagée depuis une dizaine d'années à la SNCF connaissait plusieurs limites, à commencer par la qualité des données présentes dans les gisements de l'entreprise. Le pôle données & innovation a ainsi été créé pour accélérer la transformation numérique de l'entreprise, en participant à l'industrialisation des chaînes de la donnée, sur la base de plusieurs innovations dans le domaine.
Néanmoins, la transformation numérique ne constitue pas une finalité en soi. Il s'agit en effet d'aborder la question de la mise en performance du système ferroviaire, principale promesse accompagnant la transformation numérique de SNCF réseau. En ce sens, la valorisation des données de l'entreprise constitue la deuxième composante inhérente à la création du pôle données.

À cette fin, le pôle données & innovation rassemble de nombreuses expertises qui doivent permettre de répondre à ces enjeux nouveaux. Le pôle données rassemble une trentaine de personnes, des data scientists, data engineers, des développeurs ou encore des ingénieurs ferroviaires. Après un an et demi de création, le pôle a plus d'une dizaine de projets numériques répartis au sein du groupe SNCF, essentiellement chez SNCF Réseau.

Planifier ou subir la transformation numérique

La transformation numérique doit être planifiée, au risque d'être subie. C'est un mouvement transversal à l'ensemble des activités d'une entreprise. Nous sommes tous des acteurs de la transformation numérique, dès lors que nous avons l'usage d'un smartphone ou d'un ordinateur. L'usage d'un tableau Excel ou d'une boîte e-mail ne sont pas des choses aussi anodines qu'on ne le pense. Elles participent à la numérisation des données d'une entreprise mais bien souvent sans norme ni gouvernance. Si tout cela n'est pas planifié, on aboutit à des situations non souhaitables qui peuvent être extrêmement couteuses. À titre d'exemple, la crise sanitaire a été pour beaucoup d'entreprises un accélérateur de la transformation numérique, au sens où elle a permis la dématérialisation d'échanges auparavant physiques. Pour les entreprises qui n'avaient pas les outils adaptés, cette transformation s'est bien souvent faite après avoir testé des dizaines de solutions, sans considérer les enjeux de cybersécurité. On a tous en tête ces réunions auxquelles des personnes qui n'étaient pas conviées ont accédés, y compris dans les plus hautes sphères de l'État. Outre les enjeux de cybersécurité, il faut également noter les enjeux relatifs au partage de l'information, à la définition des données, aux langages informatiques utilisés ou encore à la qualité des données.

La planification de la transformation numérique à la SNCF est récente, alors même que la transformation numérique a débuté au début des années quatre-vingt avec les premiers postes informatisés et la collecte massive d'informations.
Progressivement, la SNCF va collecter de plus en plus de données et, sur la base de ces systèmes, produire des outils qui vont participer à la mise en performance du système ferroviaire. Dans les années 2010, on dispose d'à peu près un millier d'outils et chacun dispose de sa propre base et de ses modèles de données. Cette disparité des systèmes d'informations, dont la cause est la non-planification de la transformation numérique, a créé des difficultés considérables pour faire interagir les outils entre eux, ce qui était nécessaire à la rationalisation de l'outil de production et à sa mise en performance. Ainsi, au début des années 2010, le coût de la maintenance des données pour chaque outil et des modules de conversion pour permettre les échanges entre ces derniers, ou encore celui du déploiement d'outils à l'échelle nationale sont à l'origine d'une prise de conscience et d'une réorientation stratégique.
On commence alors à imaginer autre chose et c'est à ce moment-là qu'une stratégie d'entreprise sur la transformation numérique va naître avec l'élaboration de modèles, de normes, de gouvernance des données. Cette transformation porte deux promesses :

- réduire les coûts de la transformation numérique de plusieurs centaines de millions d'euros,
- et permettre la mise en performance du système.

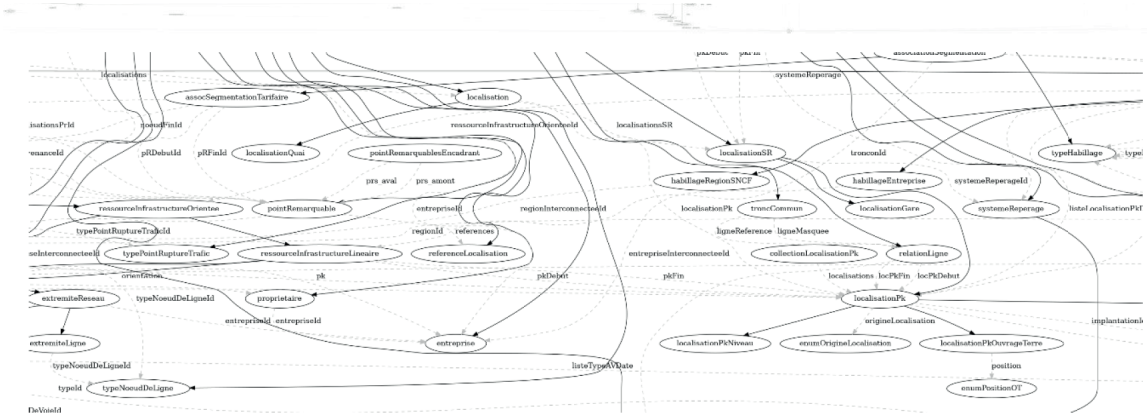
C'est dans ce contexte que le modèle d'entreprise Ariane va émerger. Il est basé sur le rail TOPOMODEL, norme internationale sur la manière de partager des données d'infrastructures entre les gestionnaires d'infrastructures, particulièrement en Europe. Il va permettre de décrire tous les objets métiers de la SNCF de la même manière, quels que soient les outils. On voit également émerger des grands gisements de données, partagés par l'ensemble des outils. On sort du paradigme « un gisement de données pour un outil ». Le grand changement de la transformation numérique n'est pas seulement de transformer les systèmes d'information mais aussi de transformer les organisations. Une organisation appelée Nouvelle Ère va également voir le jour.

Dix ans après, on constate que cette première planification n'est pas complètement réussie. Il nous reste énormément de données à numériser. On fait face à des enjeux structurants même si on arrive à avoir des possibilités que l'on n'avait pas il y a une dizaine d'années.

Les défis et le Jumeau Numérique

Premier enjeu : la qualité des données
Le principal défi auquel on fait face aujourd'hui est celui de la qualité des données présentes dans nos gisements de données. Il empêche aujourd'hui l'industrialisation d'outils à l'échelle nationale, tant l'effort de mise en qualité est important. Pour y arriver, cela suppose des données de qualité homogène, à la fois des données historiques, temps réel ou à venir pour le théorique. Cette question de la qualité de données empêche des déploiements massifs ou à des coûts très élevés. Les outils de SNCF Réseau n'ont pas totalement basculé du paradigme que j'évoquais tout à l'heure – une base de données, un outil –, vers l'usage massif de ces gisements de données. On développe, au sein du pôle Données, plusieurs solutions permettant d'accélérer cette transformation numérique, mais surtout de passer d'un paradigme de la responsabilité des données à celui de la maîtrise de données. La nouvelle ère va nommer des responsables de la donnée mais ces responsables n'auront pas les outils suffisants pour adresser la maitrise des données.

L'OPÉRABILITÉ DES SYSTÈMES D'INFORMATION



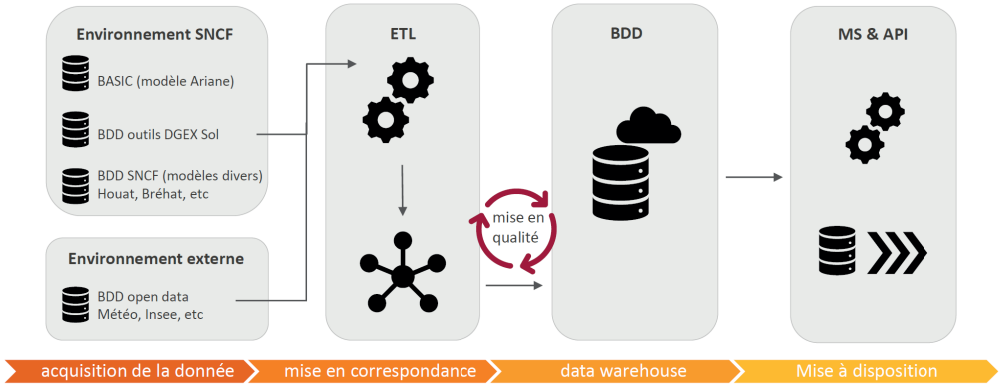
Aujourd'hui, on entre dans une nouvelle phase en essayant d'adresser la maitrise des données. C'est un des processus de mise en qualité des données.

Deuxième enjeu :
l'opérabilité des systèmes d'information
La problématique est de rendre intelligible les normes et les formats qu'on a pu créer. Il y a la planification des données qui constitue un atout mais également une difficulté pour transformer rapidement nos outils. C'est l'un des principaux freins à la transformation numérique.

En prenant l'exemple du modèle Ariane, on voit des centaines d'objets, tous reliés entre eux. C'est une manière de décrire et de modéliser l'information. C'est extrêmement complexe et lorsqu'on donne ce type de modèle à des utilisateurs pour pouvoir se saisir de la donnée, cela demande à la fois des compétences spécifiques mais aussi des connaissances métier, des connaissances informatiques pures. Aujourd'hui ces compétences sont rares mais essentielles pour accélérer cette transformation numérique. C'est l'une des problématiques de l'opérabilité mais il y en a beaucoup d'autres. Il y a celle de la performance technique des gisements, c'est-à-dire le fait d'avoir des choses « scalables », donc facilement passable à l'échelle. Hier vous aviez un gisement par outil avec des performances telles qu'elles étaient, aujourd'hui les ordres de grandeur sont totalement différents. Il y a des centaines d'outils connectés en temps réels aux gisements, qui vont consommer des centaines de giga d'informations chaque jour. Cela demande des systèmes complexes à construire et à maintenir, et l'opérabilité des systèmes d'information est l'un des grands enjeux auxquels on fait face aujourd'hui.
Troisième enjeu : La collecte des données
On a des centaines de documents dans des bibliothèques sous format pdf qui ne sont pas numérisés. Le coût de la numérisation de ces documents est considérable, on parle de saisie de données réalisées par des personnes, bien souvent des entreprises externes, pour un coût de plusieurs dizaines de millions d'euros. Quand ces données sont saisies, il faut ensuite les maintenir, c'est une autre problématique. Dans ce cadre, la reconnaissance d'image et de caractère constitue une innovation essentielle pour permettre la numérisation de ces données et son accélération. Il s'agit d'une innovation sur laquelle nous travaillons.

De nouveaux enjeux : optimiser la gestion des données
Il y a de nouveaux enjeux qui sont d'optimiser l'usage et la gestion des données.

OPTIMISER LA GESTION DES DONNÉES



Cette diapositive est une description des grandes étapes de la gestion de données aujourd'hui.
Une première étape est l'acquisition de données, internes à la SNCF mais aussi externes, notamment des données d'open source. Il s'agit des données météo, Insee, etc.
Une deuxième étape est la mise en correspondance via un ETL (Extract-transform-load). C'est un ensemble de processus qui permet notamment de structurer les données. La structuration des données est un enjeu fort, sans quoi on a des choses difficilement utilisables, en particulier dans le cadre de la valorisation des données.
Cela permet de concevoir une troisième étape, celle du stockage de l'information dans ce qu'on nomme un entrepôt de données (datawarehouse). Ce sont des données dans un mode structuré. Se pose dans cette troisième étape la question de la performance et de la construction du système. Enfin, il y a une quatrième étape essentielle qui est la mise à disposition des données, à travers ce qu'on nomme un portail API.

Mettre à disposition les données est une avancée importante dans le cadre de la planification de la transformation numérique, mais ne répond pas à l'enjeu de la rationalisation des développements au sein des outils. En effet, dans chaque outil, on va répéter des traitements de données pouvant être les mêmes d'un outil à l'autre. Ce qu'il nous faut pouvoir optimiser désormais ce n'est pas tant le partage de la donnée que les différents services que chaque outil pourra déployer dans son développement. Il s'agit de tendre vers ce qu'on nomme une architecture micro-service. Cette nouvelle problématique semble être l'un des principaux enjeux de la planification de la transformation numérique de demain. Un exemple est celui de la cartographie, utilisée par de nombreuses applications. Il s'agit de passer d'un modèle où chaque application développe sa cartographie à un modèle de service, où ce dernier est conçu une seule fois puis partagée à tous. Le service de mise à disposition des cartes constitue un micro-service sur lequel on a beaucoup travaillé.

Cela permet d'évoquer une autre construction de la transformation numérique planifiée : le Jumeau numérique. Ce dernier permet de répliquer une réalité, celle du système ferroviaire par exemple, et constitue une cible première de la transformation numérique. Outre l'intérêt quant aux données qui le compose, le jumeau numérique permet de passer à l'étape d'après qui consiste à simuler l'évolution et les dynamiques d'un système.

La valorisation des données : la promesse de la mise en performance du système ferroviaire

Une fois que la transformation numérique est faite et que l'on a des données en qualité dans les gisements, qu'en fait-on ? Il faut des endroits où innover, où être capable de construire des cas d'usage, de voir si on peut apporter des solutions face à des problématiques. C'est là qu'intervient le Datalab. Il existe depuis quelques années chez SNCF Réseau. Le datalab est un outil parmi d'autres pour permettre cette innovation. On a évoqué l'Opendata SNCF permettant d'adresser les données en dehors de la SNCF pour répondre à ces mêmes problématiques via des entreprises externes ou encore la communauté de l'open source.

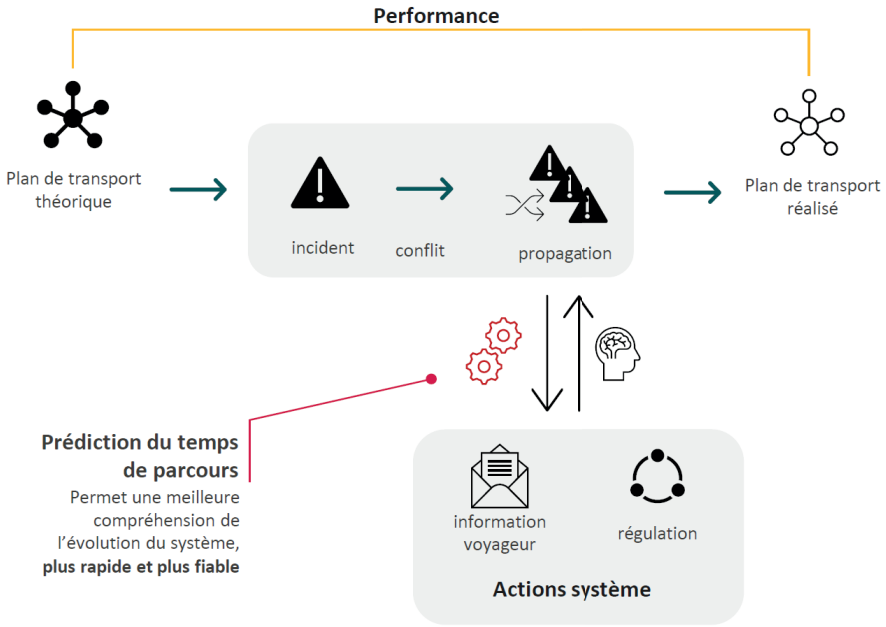
Les enjeux des datalabs

Les enjeux de l'innovation

Le Datalab a une utilité permettant de répondre aux quatre enjeux suivants :

- L'identification des cas d'usage à haute valeur ajoutée. Aujourd'hui, des plateformes nous permettent de collaborer pour partager des problématiques et d'y trouver les données, outils et information nécessaires pour imaginer des solutions nouvelles. Le datalab répond à cette identification.
- L'optimisation des temps de développement : on va avoir des temps très contraints, quelques mois ou même quelques semaines, pour montrer la capacité à réaliser des idées.
- L'intégration aux systèmes d'informations : c'est le problème de l'industrialisation. Deux tiers de l'innovation ne dépasse pas cette étape de POC (Proof of concept), de recherche et ne se transcrit jamais dans la réalité d'un milieu industriel. Le Datalab est en capacité de faciliter l'intégration dans les systèmes d'information.
- La haute disponibilité. La haute disponibilité doit permettre la disponibilité d'un service à tout moment, même en mode dégradé. On aura toujours des données disponibles même si elles doivent être en mode dégradé. On ne peut pas se permettre d'avoir des données disparaissant du jour au lendemain parce qu'un Datacenter est tombé.

POURQUOI PRÉDIRE LE TEMPS DE PARCOURS?



Un cas d'usage de datalab à la SNCF : l'usage de la prédiction du temps de parcours

Cette innovation consiste à comprendre comment le système évolue lorsqu'il est perturbé par un ou plusieurs incidents. Il s'agit notamment de comprendre la propagation des retards et des incidents dans le système. On a développé une intelligence artificielle qui, sur la base réseau neurone et d'un historique de plusieurs années de circulations, permet de comprendre la propagation du retard dans le système. Concrètement, le système permet d'estimer la position d'un train à n'importe quel moment de sa circulation, de comprendre les interactions dans le système, et notamment les éventuels conflits.

En conclusion, l'usage qu'on a des datalabs nous permet de faire le lien entre la transformation numérique et la valorisation des données. C'est une passerelle entre ces deux mondes.

Brigitte GUIGOU
Je vous remercie pour cette intervention qui nous permet de mieux comprendre comment une entreprise comme la SNCF fait face à la fois à la question de la transformation numérique et à celle de la mise en performance du système ferroviaire. Pour revenir sur la question de l'innovation, qu'est-ce qui vous a permis d'innover ? Quels sont les leviers et prérequis ?

Guillaume LECOEUR
Il faut d'abord accéder à des données en qualité. Pour mettre en place cette intelligence artificielle, il a fallu trois ans de données avec un historique présentant la position des trains sur tous les parcours. On parle de centaines de giga de données. Il nous a fallu accéder à ces données de manière performante pour industrialiser la solution. C'est l'un des premiers enjeux. Un deuxième enjeu est de développer des intelligences artificielles et donc d'acquérir ces compétences nouvelles dans l'entreprise. Le troisième enjeu est celui de l'industrialisation. Une fois qu'on a eu quelque chose qui fonctionnait sur nos environnements, il a fallu passer à l'échelle de manière rapide. On arrive à adresser cette problématique dans le cadre du Datalab, passer de l'état de la preuve du concept à l'état industriel.

Brigitte GUIGOU
Combien de temps vous a-t-il fallu pour mener à bien cette innovation ?

Guillaume LECOEUR
Cela a été fait en moins de 6 mois. Le processus est d'ailleurs toujours en cours. On est dans la phase d'industrialisation actuellement. Il faudra moins d'un an pour concevoir l'idée depuis le cas d'usage jusqu'à l'industrialisation. On espère une industrialisation au mois de juin 2021.

Brigitte GUIGOU
Comment avez-vous travaillé avec les experts métier ? Avez-vous des modalités de travail spécifiques ?

Guillaume LECOEUR
L'équipe rassemble des expertises variées. Les data scientists ont de très bonnes compétences dans le domaine de la data mais ne connaissent pas nécessairement bien les métiers ferroviaires. C'est pourquoi il est essentiel d'assembler dans une même équipe des compétences diverses et les faire discuter. Il faut trouver un langage commun pour faire émerger des solutions pertinentes. Faire émerger une organisation mixant ces compétences dans une même équipe est l'un des principaux éléments de la réussite de ces projets.

Brigitte GUIGOU
Et cette collaboration fonctionne ?

Guillaume LECOEUR
Elle fonctionne. Elle sort des habitudes de travail classique à la SNCF et permet des collaborations nouvelles, avec un ancrage avec le métier très fort. En ce sens, la donnée est un actif réellement partagé. Elle n'est pas seulement la ressource de travail de quelques personnes dans l'entreprise, elle est partagée et accessible à tous.

Brigitte GUIGOU
Avez-vous rencontré des réticences en interne au groupe SNCF sur cet enjeu de transversalité ? Quels sont, d'une façon générale, les principaux obstacles auxquels vous êtes confrontés ?

Guillaume LECOEUR
On fait toujours face au défi de la transformation numérique, dont on ne peut pas encore dire qu'elle est pleinement réussie. Les défis sont tout à la fois techniques et organisationnels. De même, on fait également face à d'autres entités portant l'innovation dans l'entreprise avec des modes de fonctionnement différents, ce qui peut parfois créer des réticences mais qui est aussi la garantie d'avoir plusieurs pensées dans l'entreprise, qui peuvent être confrontés dans des lieux de débat que sont les datalab.

Antoine COURMONT
La présentation de SNCF Réseau est passionnante. Guillaume Lecoeur a bien souligné les défis pour mettre en place ce genre de structure. Toutes les entreprises engagées dans la transformation digitale dans le secteur urbain mettent d'ailleurs en place le même genre de dispositif, que ce soit en matière d'entrepôt de données puis de transformation en interne de l'usage de ces données.
J'ai une question, vous avez initié un partenariat avec Open Street Map. Avez-vous d'autres associations avec ce genre de producteurs de données issues de la société civile ?

Guillaume LECOEUR
L'exemple d'Open Street Map est le plus significatif. Il constitue tout à la fois une source de données et une manière de la maintenir. On a aujourd'hui une connaissance partielle du patrimoine des gares et de l'accessibilité aux gares. On se base sur la donnée Open Street Map pour une première ébauche de ces données et ensuite pour les mettre à disposition. Le problème est de pouvoir contrôler et maitriser l'évolution des données une fois qu'elles sont mises en qualité et distribuées dans Open Street Map. C'est une des collaborations qu'on peut avoir avec l'environnement Open Source, notamment sur la question des formats. On utilise l'Open Data de manière globale, telles que celles de l'Insee par exemple.

Brigitte GUIGOU
Je vous remercie et vous invite à consulter l'ensemble des ressources de ce petit déjeuner – vidéo, diaporama et synthèse – sur le site web L'Institut Paris Region.

L'INSTITUT PARIS REGION
EST UNE ASSOCIATION LOI DE 1901

15, RUE FALGUIÈRE - 75740 PARIS CEDEX 15 - TÉL. : 01 77 49 77 49