ARTICLE DU DOSSIER

« Mettre l'utilisateur au centre de la diffusion de données »

Entretien avec

Christian QUEST

Coordinateur de la Base Adresse Nationale au sein de la mission Etalab, président d'Open Street Map d'avril 2014 à juin 2017[1].

La loi pour une République numérique a impulsé la création d'un service public de la donnée, dont une des missions centrales est de mettre à disposition des bases de données dite de référence. Ces données « constituent une référence commune pour nommer ou identifier des produits, des services, des territoires ou des personnes ; sont réutilisées fréquemment par des personnes publiques ou privées autres que l'administration qui les détient ; leur réutilisation nécessite qu'elles soient mises à disposition avec un niveau élevé de qualité ». La base adresse nationale est l'une des neuf bases de données de référence. Dans cet entretien, nous revenons avec Christian Quest, coordinateur de la base adresse nationale au sein de la mission Etalab et président d'OpenStreetMap France de 2014 à 2017, sur l'origine de cette base et les conditions nécessaires pour que celle-ci devienne une donnée de référence.

AST[2] : Peux-tu nous présenter le projet de la base adresses nationale ?

CQ : Avant de parler des bases de données adresses, le premier problème concerne l'existence même des adresses. En France, l'adresse est gérée au niveau des communes : on en a maintenant un peu moins de 36 000, mais une très grande majorité sont toutes petites et ne sont pas outillées pour gérer des adresses, voire même n'ont jamais donné de nom de rue ni de numéro. La Poste indique que 40 % des points d'arrêts postaux, c'est à dire l'endroit où le facteur s'arrête pour distribuer du courrier, n'ont pas de numéro. Il s'agit de lieux-dits ou des hameaux où il n'y a pas d'adresses. Les maires n'ont pas d'obligation de dénommer les voies ni de les numéroter. Ils le font parce que c'est nécessaire pour la bonne organisation de la commune, pour les secours, etc. Mais, il n'y a rien qui impose l'adressage aujourd'hui. C'est un premier problème que la base adresse nationale ne va pas résoudre : une base adresse, ça ne peut répertorier que ce que l'on a nommé et numéroté.
Le deuxième problème n'est pas que l'on n'ait pas de base adresses nationale, c'est qu'il en existe plusieurs : le cadastre, la BD Adresses de l'IGN, le Répertoire des Immeubles localisés (RIL) de l'INSEE, les bases de données de La Poste, sans compter toutes les entreprises (GRDF, Enedis, Orange, etc.) qui ont des bases adresses. Beaucoup d'acteurs se sont créés des bases pour leurs propres besoins. Mais, ces bases adresses ne contiennent pas les mêmes informations. La Poste, ils ont 18 ou 19 millions d'adresses, l'IGN, 25 millions, moi, j'estime

1. OpenStreetMap (OSM) est un projet collaboratif visant à constituer une base de données géographiques libre.
2. Entretien réalisé le 21 avril 2017 par Antoine Courmont, Samuel Goëta et Timothée Gidoin (« AST »)

qu'il y en a environ 20 millions d'adresses sur le terrain en France. Finalement le problème des bases adresses en France, ce n'est pas qu'il n'y en a pas, c'est qu'il y en a trop et qu'il n'y en a aucune qui soit arrivée à un niveau de qualité qui fasse que l'on ne se pose même plus la question de savoir laquelle on utilise.

AST : Quand est venu le projet d'unifier ces différentes bases de données ?

CQ : Cela a commencé au sein d'OpenStreetMap. Nous étions régulièrement sollicités au sujet des adresses, or, nous n'avions aucune base de ce type. En même temps, nous commencions à en avoir besoin car de nombreuses données en *open data* contiennent des adresses sémantiques, mais aucune localisation géographique. Or, si nous souhaitons les remettre sur nos cartes pour ajouter par exemple les monuments historiques, il nous fallait pouvoir géocoder ces adresses sémantiques et obtenir une position géographique. Un jour, un de nos contributeurs a écrit un script qui nous a permis de récupérer les adresses à partir des plans cadastraux, et ainsi de générer une base de 16 millions d'adresses. Nous avons croisé cela avec les adresses existantes dans OSM et celles publiées en open data par certaines collectivités locales. Cela nous a permis de créer une base qui prenait le meilleur des trois que l'on a décidé d'appeler la Base Adresse Nationale Ouverte (BANO). Assez rapidement, à partir de la BANO, qui est en fait un outil de diffusion des données collectées dans OSM, on s'est dit qu'il faudrait conclure des partenariats nationaux, ou locaux, pour alimenter, mettre à jour, agréger et verser un maximum de données de qualité. Notre souhait était de mettre en place un pot commun où tout le monde vient mettre de l'adresse et tous ceux qui en ont besoin viennent se servir. C'est l'idée d'OSM, mais appliquée à une thématique unique qui est l'adresse. En faisant cela, nous avons donné un grand coup de pied dans la fourmilière : cela faisait des années que l'on entendait parler d'une base adresses nationale et qu'elle n'existait pas dans les faits.

AST : Comment expliques-tu qu'un projet qui était assez historique, d'une base adresse nationale qui rapproche les différentes bases des différentes administrations, n'ait jamais pu voir le jour, et que ce soit le côté ouvert qui ait permis d'unifier un peu tout ça ?

CQ : C'est parce qu'on est dans une logique ouverte justement. Ça dépend où on met la priorité : est-ce qu'on met la priorité sur l'intérêt des utilisateurs des données, ou est-ce qu'on met la priorité sur les producteurs des données ? Avec BANO, on a mis l'intérêt sur les utilisateurs. Ils ont besoin de quelque chose, on a de quoi leur répondre, on fait. Nous n'avons pas de contraintes financières ou de business model à défendre. Lorsque c'est le cas, malheureusement, l'intérêt du producteur tend à passer souvent devant l'intérêt des utilisateurs. Et c'est ça qui bloque la mécanique.

AST :A partir de la BANO, comment est-on arrivé au projet de la base adresse nationale (BAN) impulsée par Etalab ?

CQ : La BANO a fortement intéressé Etalab. Ça faisait des années qu'ils avaient un problème tout bête : plein de jeux de données en open data avec des adresses, et pas de possibilités simples de les géocoder, sauf en utilisant les services de Google ou similaires, qui posent d'énormes problèmes juridiques et de souveraineté. Si un Etat n'est pas capable de faire ce genre de chose soit même sans faire appel à une multinationale étrangère, c'est un peu gênant. Donc, ça a fortement intéressé Etalab, et ils m'ont recruté pour faire avancer le dossier adresse le plus loin possible.

AST : Pourquoi avoir conclu un partenariat entre l'IGN, La Poste et OSM en avril 2015 ?

CQ : En mettant en place un partenariat avec les principaux acteurs producteurs de bases adresses en particulier La Poste et l'IGN, l'objectif de la BAN est de dénombrer toutes les

adresses, savoir qu'une adresse existe. Il faut croiser et agréger ces bases de données. Ce n'est pas facile à cause du manque de circulation de la donnée entre les administrations et les différents services de l'État. Obtenir 90% des adresses, c'est assez facile. Faire les quelques % restants, c'est plus compliqué. Si chacun fait le boulot dans son silo, il parviendra à couvrir 90 voire 95%. Si par contre on travaille tous en commun, on va tous amener l'énergie qui va permettre d'atteindre les 100%.

AST : Et à l'heure actuelle, comment se fait le mélange entre les données de La Poste, les données d'OSM et les données IGN ?

CQ : Alors pour l'instant, la BAN qui est diffusée aujourd'hui, la BAN que moi j'appelle version 0, ce sont des données qui sortent de l'IGN. L'IGN agrège des données qui proviennent du cadastre, des données qui proviennent de La Poste, des données qui proviennent d'un certain nombre de partenaires, de quelques collectivités, de quelques services de pompiers, de gens qui collectent aussi des informations pour l'IGN. Il n'y a pas de données OpenStreetMap, pour une simple raison, c'est qu'il y a une incompatibilité de licence. Les données OpenStreetMap sont sous licence ODbL[3], qui implique un partage à l'identique, alors que l'IGN va commercialiser ses données sous une licence qui n'implique pas le partage à l'identique. Le problème n'est pas la commercialisation : on peut commercialiser des données OpenStreetMap, mais, même si on les commercialise, il faut les commercialiser avec une clause de partage à l'identique. On ne peut pas s'en séparer. Et cette clause n'existant pas, il est impossible de diffuser de la donnée provenant d'OpenStreetMap. Donc la BD adresses ne peut pas intégrer des données OpenStreetMap.

AST : Il y a donc deux bases adresses nationales : la BAN et la BANO ?

CQ : Oui, il reste deux bases. Alors, au niveau d'OpenStreetMap, on a aussi fait le choix de ne pas faire un cumul entre BAN et BANO. Si on faisait le mix des deux, on aurait, une base qui serait plus à jour et plus complète que la BAN, et si OpenStreetMap fait ça, en gros on tue la BAN parce que les utilisateurs attendent la base la plus complète et la plus à jour. On s'est un peu interdit de le faire pour laisser une chance à la BAN de prendre, en visant le long terme quitte à déroger à notre règle "je peux faire, je fait".

AST : Qu'est-ce qui a donc changé avec la BAN ?

CQ : Ce qui a changé avec cette convention, c'est qu'il existe une base adresses de sources officielles (IGN, La Poste), disponible sous une licence de type ODbL, qui répond aux critères de l'open data. Cela a permis à Etalab de mettre en place un géocodeur qui a bonne réputation. En 2016, 446 millions de requêtes ont été traitées, en 2017 plus d'un milliard de requêtes. Par exemple, depuis que la base SIRENE est en open data, tous les mois on a un stock mensuel qui sort, et, tous les mois, je le géocode. Beaucoup de gens sont déjà extrêmement contents et surpris qu'on fasse ça. Ça leur évite de le faire, parce que c'est quand même un boulot qui est un peu compliqué, et qui nécessite des ressources et des compétences.

AST : Quelles sont les prochaines étapes pour la BAN ?

Il faut que l'on arrive à passer à une vraie gestion collaborative des adresses. Quand je parle de collaboratif, ce n'est pas uniquement entre les signataires de la convention, c'est plus largement collaboratif avec la DGFIP et le cadastre, avec un maximum de collectivités, avec un maximum d'autres opérateurs, avec l'INSEE. Pourquoi aujourd'hui l'INSEE gère-t-elle sa propre base

3. L'Open Database License (ODbL) est une licence de style « copyleft » permettant de copier, de modifier, de faire un usage commercial sous trois conditions : citer la source, redistribuer sous des conditions de partage identiques les modifications, maintenir ouverte la base de données redistribuée.

adresses ? Je leur ai posé la question : la qualité. C'est pour des raisons de qualité qu'ils ont leurs propres données. Cela signifie que le produit disponible aujourd'hui ne convient pas. Donc, plutôt que de travailler tout seul de son côté pour maintenir les données, il faut collaborer avec les autres acteurs sur le même domaine. Ces projets coopératifs, c'est une forme de coopérative de fainéants. Si on additionne notre travail plutôt que de refaire la même chose chacun de notre côté, soit on fait globalement moins de travail, soit on obtient un meilleur résultat avec la même quantité de travail. C'est ça que j'essaye de faire comprendre. Mais ce n'est pas évident parce que la culture de la collaboration, du partage de l'information, du partage des données est très limitée voire inexistante. La culture dominante, aujourd'hui, c'est l'inverse. C'est vraiment très très compliqué de faire comprendre que si on veut arriver à fournir le meilleur service final à l'usager, et je reviens à l'usager, il faut qu'on s'allie. Ensemble, on est plus fort pour abattre les derniers problèmes.

AST : C'est cela qui permettra à la BAN de devenir véritablement une « donnée de référence »[4] ?

CQ : Pour qu'une base fasse référence, en fait il faut que qu'elle soit la plus fraîche possible. La fraîcheur des données est presque plus importante que le critère certifié d'autorité ou de qualité de la donnée. C'est vraiment la fraîcheur. Si la longueur du tuyau entre l'apparition d'une donnée sur le terrain et la sortie dans le référentiel est trop longue, qu'est-ce que font les gens ? Ils font une copie du référentiel et puis ils font les mises à jour eux-mêmes. Finalement c'est pour ça que les pompiers ont des bases adresses. Ils ne devraient pas en avoir, ils devraient récupérer la base officielle et puis terminé. Pourquoi les pompiers mettent à jour leur base adresses ? Il y a un truc qui ne va pas. Vraiment la fraîcheur de la donnée, c'est un critère primordial pour devenir une référence. Une référence ça ne s'impose pas. C'est exactement comme un standard, les gens vont converger vers quelque chose en disant, "on sait que ça marche, c'est bon, et on ne se pose pas la question". On utilise parce que l'on sait qu'on ne va pas trouver ou faire mieux.

Pour arriver à cela, il faut mettre la priorité sur l'utilisateur et non sur le producteur. Par exemple, l'INSEE est utilisateur de ses propres données, donc ils ont une logique d'utilisateur, ils produisent des données parce qu'ils les utilisent eux-mêmes. Ok, très bien, ça c'est une chose. Mais après, quand on a des utilisateurs extérieurs, il faut aussi se mettre à leur place. Dans un domaine tout autre, qui n'est pas l'adresse, mais qui est quand même indirectement lié à l'adresse, la gestion du code officiel géographique est aujourd'hui totalement inadaptée aux utilisateurs. Les changements qui interviennent dans le découpage administratif français, ce n'est pas six ou dans le meilleur des cas trois mois après leur entrée en vigueur, que l'information doit être diffusée ! C'est trois mois AVANT l'entrée en vigueur qu'elles devraient être diffusées pour que les utilisateurs puissent anticiper ces changements. À tel point que même en interne à l'INSEE, sur le code officiel géographique, il y a un code officiel géographique officieux, temporaire, qui est utilisé sans attendre le mois de mars ou avril, le code officiel officiellement publié, parce que sinon, pendant trois mois, on ne pourrait enregistrer aucune nouvelle entreprise dans la base SIRENE. En interne, il y a un Code Officieux Géographique qui est produit parce qu'ils en ont besoin sans attendre, mais les ré-utilisateurs à l'extérieur qui ont exactement les mêmes besoins attendent des mois. C'est un vrai problème. Pour certaines données de référence, il faudrait peut-être dissocier le rôle de producteur du rôle de premier utilisateur de la donnée. Sinon, il y a une logique métier dans la production des données, or, une donnée de référence ne doit pas être faite pour un métier, elle doit être faite pour tous les métiers. Il faut donc vraiment, vraiment, mettre l'usager de la donnée au centre. Et le premier usager de toutes les données de l'État, c'est l'État lui-même. Or, aujourd'hui, il y a une très mauvaise circulation des données. On s'aperçoit que le travail a été refait plusieurs fois parce qu'on ne s'échange pas les données.

4. Au sens de la loi de 2016 ; voir : https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000031366350&idArticle=LEGIARTI000033205649&dateTexte=&categorieLien=cid