OPEN DATA: LES DONNÉES LIBÉRÉES DOIVENT-ELLES GRATUITES?

Introduction
Retours d'enquêtes
Script de l'objet Donnée ouverte.
Format du conflit et acteurs impliqués.

Introduction

Pour avoir une idée convenable de ce qu'est l'open data commençons par définir ce qu'est une "donnée"
La donnée comme on l'entend ici est un fait auquel on apporte aucune valeur ajoutée, par exemple: "le 23 décembre 1995 à paris il faisait 18 degrés" est une donnée, alors que "le 23 décembre 1995 à paris il faisait 18 degrés, ce qui est très chaud pour un hiver" n'est pas une donnée car on interprète la donnée. Une manière plus simple de bien différencier donnée et information est de faire une analogie entre un fait et le commentaire de ce fait.
Maintenant passons à "donnée ouverte":
"Une donnée ouverte est une donnée publique brute, qui a vocation à être librement accessible et réutilisable. La philosophie pratique des données ouverte préconise une libre disponibilité pour tous et chacun, sans restriction de copyright, brevet ou autre mécanisme de contrôle"
Il est à noter qu'il n'y a pas de définition formelle de ce qu'est vraiment une "donnée ouverte", car celle-ci varie selon les interlocuteurs et les pays, cependant les considérations d'accès libre et gratuit à des données à jour, sous format accessible et non propriétaire (pour permettre un traitement informatique automatique) sont des constantes de ces différentes définitions
Tant que nous sommes aux définitions profitons-en pour définir ce qu'est une "donnée publique", à ne pas confondre avec une donnée ouverte ce qui pourrait arriver au lecteur inaverti:
une donnée publique est une donnée produite (i.e récoltée et publiée) par l'état une collectivité régionale ou un organe parapublic dans le cadre d'une mission de service public.
Dans le cadre de notre étude nous nous limiterons au cadre français (les lois varient énormément d'un pays à l'autre, il ne serait alors pas pertinent de faire une analyse à ce sujet). Cependant nous nous permetterons de parler à de nombreuses occasions du mouvement Open Data anglais, car il est à la fois l'un des pionniers du domaines et un de ceux qui s'est le mieux implanté.
Ainsi, continuons avec la présentation des retours d'enquêtes pour installer un peu de rigueur dans notre approche du problème !

Retour d'enquêtes

Les textes de loi:

Le premier texte de loi traitant de l'ouverture des données publiques est la loi du 17 juillet 1978, elle stipule qui doit libérer un certain type de données et les conditions de ces libérations. Concrètement toutes les données (sauf celles relatives à la défense et autres sujets sensibles, personnelles et grénérées par les EPIC) doivent être libérées et le coup de leur revente ne doit pas excéder le coup de leur création.

Vient ensuite le décret 2005-1775 qui précise les cas de litige et introduit la CADA (l'institution qui décrète si une demande de libération est recevable ou non), il permet également de refuser de libérer des données personnelles si le coût de leur anonymisation est jugé trop élevé.

Circulaire du 26 mai 2011; cette dernière donne les grandes lignes de la politique de l'état face au mouvement open data et donne des indications quand aux formats des données. Elle indique également que seule certaines données peuvent donner lieu à une redevance (et non plus l'inverse, c'est à dire que seules certaines données doivent être gratuites).

Interviews, sitographie et revue de presse

Maintenant que le tout est bien cadré, nous allons pouvoir commencer à vraiment nous intéresser à la controverse liée au mouvement open data français.

Concrètement la question qui revient le plus souvent est celle des coûts et des bénéfices liés à la libération des données, en effet il est évident que, à de rares occasions (par exemple celles où ces données peuvent porter atteinte à l'état) personne n'est opposé à l'open data. En effet c'est un mouvement qui crée de l'innovation et de l'emploi dans la mesure ou de nombreuses start up et entreprises utilisent ces données pour créer de la valeur ajoutée (des applications mobiles ou des algorithmes de traitement des données). Ainsi, il semblerait qu'à première vue tout le monde y gagne, même l'état car cela crée de l'emploi, se pose alors la question "Mais où est le problème, pourquoi toutes les données ne sont elles pas déjà libérées ?"

L'un des aspects les plus important dans l'open data est la facilité d'utilisation des données par les ordinateurs, et c'est là que tous les problèmes se posent, en effet lors de leur récolte et création ces données ne sont presque jamais rentrées sous un format adapté à cela (par exemple sous word ou sous adobe au lieu de libreOffice), ainsi leur mise à disposition sous ces formats requiert un travail supplémentaire, et en conséquence un coût supplémentaire qui n'était pas présent avant et qui ne rentre pas forcément dans les budgets des actants concernés, de même ces données sont parfois immédiatement tranformées en informations il faut alors les "nettoyer" pour pouvoir les libérer (par exemple, si un département a un budget de 5k€, que sa mission requiert ces 5k€ et que la mise en forme des données coûte 1k€ il ne peut faire ni l'un ni l'autre). Ainsi la mise à disposition gratuite de ces donnnées est très contraignante à court terme, il paraît ainsi cohérent de faire payer ce travail supplémentaire engendré par une demande qui n'était pas là auparavant. Certains feront remarquer que ces données devant être créées dans le cadre de mission de service publique, elles appartiennent au domaine public et que leur utilisation devrait être gratuite, car déjà payée par le contribuable. Il faut également remarquer que dans le contexte économique actuel cette question de budget se fait d'autant

plus ressentir.
Un exemple assez marquant de cette dualité entre envie d'avoir des données et réalité de leur coût est le cas du site nosdeputes.fr qui permet de voir les différentes activités des députés, cette initiative est actuellement payée de la poche des dits députés, cependant que se passerait-il si ce même site générait des bénéfices via l'ajout de publicité ? Certes ces dernières couvriraient des dépenses (réelles) mais le contribuable se sentirait trompé ("Comment se fait-il que mes députés se fassent de l'argent sur mon dos alors qu'ils font juste leur boulot?"). Ainsi on voit bien le problème lié au fait de vendre des données publiques, alors même qu'elles peuvent être déjà considérées comme payées par les impots.
D'un autre côté cette monétisation des données peut empêcher la création de certaines entreprises qui ne peuvent lever assez de fonds pour payer ces données, ce qui est regrettable car ces dernières génèrent de la valeur, et bien souvent rapportent plus en impots que le coût de création des données qu'elles utilisent.

Comme nous avons pu le voir avec les textes de loi, les données doivent être libérées si elles sont demandées, si ce n'est pas le cas l'acteur concerné a le droit de saisir la CADA, puis si cette dernière l'y autorise, à saisir la justice pour avoir accès à ces données. Cepandant dans les faits de nombreuses données ne sont pas libérées et la CADA est de plus en plus souvent saisie, de même les données actuellement libérées sur data.gouv.fr sont souvent sans grand intérêt (du type cartes géographiques). Ainsi l'Etat ne respecte pas vraiment ses propres lois dans les faits, car toutes les données ne sont pas libérées mais uniquement celles qui ont un faible coût de libération. De même dans le cas des EPIC tels l'IGN ou la RATP la revente de ces données est le but principal de ces dernières et leur seul moyen de rentabilité.

Une des raisons principale du peu d'enclin des collectivités territoriales et de l'état est le manque actuel de structure économique prête à accueillir et utiliser ces données. En effet les acteurs de l'open data souhaiteraient un modèle qualifié de Data Driven, c'est à dire de commencer par libérer les données puis de déterminer après leur libération comment les utiliser, à l'instar de DBpedia.org. L'état ne pouvant pas être sûr de la rentabilité de ses coûts n'est donc pas forcément ravi de devoir investir sans avoir la moindre garantie derrière. Plus généralement le mouvement Open Data français est relativement récent (le portail data.gouv.fr est arrivé dans les derniers, après celui d'Angleterre et des Etats-Unis) et se cherche encore, bien que les acteurs pricipaux tels liberTIC, etelab
data publica soient déjà présents.

Le problème central de notre controverse soulevé par ces retours est donc celui du coût de fabrication des données, en effet ce coût est demandé par les acteurs d'être assumé par un organisme public qui ne voit pas de bénéfices rentrer malgrés des frais consistants. Est-ce à l'Etat d'assumer ces coûts,, et si non, qui doit les assumer ?
Les problèmes secondaires et sous-jacents sont ceux de l'anonymisation des données (cette dernière revient cher et est souvent remise en question), de la mission de l'état envers les entreprises et contribuables et le libre accès à l'économie (au sens de pouvoir créer une entreprise).

Script de l’objet Donnée ouverte

L’objet qui nous intéresse ici est l’objet de la donnée ouverte.
Commençons par nous intéresser au moment de l’apparition de cet objet. On le voit pour la première fois vraiment apparaître avec l’apparition de dbPedia (qui est un précurseur du mouvement) puis du portail data.gov.uk et de data.gouv.fr. Cet objet est donc relativement récent (2007) par rapport à l’arrivé de l’Internet (1989-1990) et on peut noter que les deux ont pour créateur (ou cocréateur pour le premier) Tim Berners-Lee. C’est également autours de ces dates (200*) que les ordinateurs avec des puissances de calcul suffisante pour traiter de grosse bases de données se sont répandus. De même le mouvement Big Data s’est également formé à ces dates là.
Une raison plausible de l’apparition de la donnée ouverte est le regain d’intéret pour la donnée brute. En effet avec l’Internet, la quantité d’information disponible a fait une croissance exponentielle, cependant l’information n’est pas (encore) utilisable par les ordinateurs, de même bien que nombreuses elles sont rarement réunnies et sont souvent issues de différentes études ou jeux de données (ce qui leur fait perdre toute valeur d’un point de vue technique). Ainsi dans le cadre des études scientifiques et des entreprises le besoin de données brutes et issue d’une même étude avec le même protocole s’est fait de plus en plus ressentir (d’autant qu’elles possèdent enfin les capacités de calcul permettant de traiter autant de données).
Nénmoins de tels jeux de données ne peuvent être générées par les entreprises (elles le feraient payer le prix fort, ce qui en rebouterait plus d’un et laisserait le marché stagnant dans le meilleur des cas, d’autant qu’elles ne peuvent être forcée à libérer leurs données), c’est alors vers l’état que ces dernières se sont tournées. En effet il mène de nombreuses missions à même de générer des données pertinentes et utiles, et il peut également être fortement incité à les libérer (après tout ce sont des impots de ces mêmes entreprises que ces missions sont payées). Ces données étant publiques une pression a pu être mise en place de la part des associations et entreprises quand à la qualité, la quantité et la gratuité de ces données. D’autant que leur usage pouvant créer de l’innovation et de la valeur ajoutée, l’Etat avait des raisons de libérer ses données.
Actuellement que l’objet Open Data a été relativement bien implanté les données sont libérées, mais la question du coût de la libération/création/structuration se pose de plus en plus, au fur et à mesure que ces dernières sont libérées. On a ainsi vu les dernières lois sur le domaine permettre de nouvelles raison de refuser de libérer ces données (coût d’anonymisation jugé trop élevé, ou alotrs présence dans un base de données). C’est ainsi que la controverse autours du coût des données libérées est apparu.
Aux Etas-Unis, le script a été modifié par différents groupes (tels la Sunlight Foundation) qui se sont saisi du mouvement pour en faire un outil de surveillance civile et citoyenne des activités de l’état.

Format du conflit et acteurs impliqués.

La controverse du coût des données ouverte n'est pas vraiment un conflit dans la mesure où il n'y a pas de réelles oppositions de convictions ou de croyances mais de simple divergences sur la manière de s'yprendre, et dans le pire des cas une forte réticence.
La controverse se rapproche cependant dans sa forme à un débat car il y a de nombreuses associations impliquées ainsi que de nombreux journaux grands public (the Guardian, the Economist, Le Monde …).
L'arêne de ce débat est le plus souvent la presse (écrite et en ligne) qu'utilisent les association pour communiquer auprès de l'Etat et des collectivité locale, dans le cas de désaccord entre les acteurs public et un autre acteur, ce dernier peut saisir la Cada et, si elle l'autorise, cette arêne se déplace au tribunal.
Ainsi de nombreux acteurs sont des associations (liberTIC et quadrature du net) ou des structures liées à l'Etat: les collectivitées territoriales (la ville de Rennes en est un des acteurs les plus actifs), le portail Etalab, des Autorités Administratives Indépendantes (CADA et CNIL), EPIC (RATP, SNCF et IGN). Il y a également des Start-up et autres entreprises dont le but est l'utilisation de ces données.
Ces trois différents types d'acteurs ont chacun un rôle différents, en effet:
- les acteurs publics ont une place plus passive dans le débat, ils le subissent en quelque sorte car c'est de leur part qu'est exigée la production de données, cependant il y a aussi quelques acteurs plus "actifs" : Etalab et la Cada. Ces deux derniers libèrent les données et déterminent la politique de libération pour le premier, et détreminent la recevabilité des plaintes à propos de la libération des données. Actuellement ces acteurs tendent à limiter l'importance du mouvement open data tant qu'un modèle économique viable n'a pas encore été mis en place.
- les acteurs associatifs sont relativement en accord avec les acteurs privés (entreprises), tous deux veulent une libération gratuite des données pour faciliter l'innovation (ou tout simplement économiser de l'argent), la création d'entreprises et d'emplois.
Nous pouvons également que parmis ces acteurs très peu peuvent être qualifiés d'experts technique (à la limite la Cada, le Cnnum et le CNIL).