PRESENTATION

Appel à projet : « Développement de l'open data au sein des territoires »

Programme : Industrialisation de la mise à disposition de données ouvertes

Investissements d'avenir -
Transition numérique de l'État et modernisation de l'action publique

PLATEFORME NATIONALE DE QUALIFICATION DES DONNNEES LOCALES (QUALIDATA)

SYNTHESE DU PROJET

Le développement de l'ouverture des données publiques, engagé en France depuis près de sept ans, prend aujourd'hui une dimension incontournable dans les stratégies numériques des territoires, de l'Etat et de nombreux acteurs privés. La Loi PRN sanctuarise le principe d'ouverture des données par défaut et de très nombreuses collectivités de petite et moyenne taille sont dorénavant concernées.

Le passage à l'échelle de l'ouverture des données présente un défi important : une plus grande quantité de données sera publiée mais ***la qualité sera-t-elle au rendez-vous ?***

Dès à présent, la description, les formats et le contenu des données publiées par les collectivités pionnières de l'opendata pose un problème de consistance. Plus de 20 000 jeux de données sont ouverts par plus de 200 collectivités de toute taille et de tout niveau hiérarchique. Si le mouvement est remarquable et très mobilisateur, le manque de cohérence entre ces données engendre un gisement national de données difficile à exploiter et à comprendre. Cela fait peser sur l'opendata des contraintes importantes pour sa réutilisation et peut véritablement mettre en question sa crédibilité.

Le gouvernement a anticipé ces difficultés en confiant à Opendatafrance la mise en œuvre de dispositifs d'accompagnement des collectivités pour favoriser la normalisation et la production des données produites au niveau local. Ce projet appelé, OpenDataLocale, agit en quelque sorte pour structurer des gisements locaux « multiformes » en briques de données interopérables et cohérentes.

Comment s'assurer que ces données sont dans un état de structuration facilitant au maximum leur réutilisation ? Comment s'assurer qu'elles sont conformes aux recommandations et aux standardisations ?

Cela vaut pour les collectivités qui souhaitent vérifier que leurs données correspondent bien aux recommandations : un outil de test et d'assurance qualité leur est nécessaire.

Cela vaut aussi pour les acteurs nationaux qui s'intéressent aux données locales dans leur ensemble, que ce soit pour constituer des référentiels nationaux (Data.gouv.fr) ou pour des besoins d'analyses (observatoires, datascientists, entreprises…) : ils doivent se « nourrir » de centaines, demain de milliers de données issues des niveaux locaux et l'acquisition d'information de mauvaise qualité détériorera la base globale des données traitées, ruinant les résultats ou demandant des efforts considérables de retraitement.

Les partenaires du présent projet proposent de mettre en place une plateforme nationale capable ***de tester la qualité des données publiées au niveau local****, de qualifier les sources suivants plusieurs niveaux d'exigence et lorsque les données sont suffisamment fiables,* ***de permettre des exploitations cohérentes au niveau national****. Elle complète les dispositifs déjà existants sur data.gouv.fr et se positionne comme une couche dédiée à la qualification et le pré-traitement des données des territoires.*

Cette plateforme a pour missions (finalités) :

- *d'aider les collectivités à valider la cohérence de leurs publications et d'améliorer les données qu'elles publient grâce à des comptes-rendus de tests détaillés;*
- *de faciliter le travail des collectivités sur la qualité de leurs données, enjeu majeur de leurs projets ainsi que de la mutation de leurs systèmes d'information ;*
- *de certifier les sources (un gisement de données plutôt qu'un portail dans son ensemble) ayant un niveau de qualité satisfaisant;*
- *de nourrir des référentiels permanents (ex data.gouv.fr) ou temporaires (une analyse nationale sur l'ensemble de jeux de données d'une thématique)*
- *de fournir des indicateurs qualifiés en vue de l'évaluation de la situation des opendata dans les territoires (observatoire dédié ou injecteur pour des observatoires-tiers)*

Les impacts attendus sont :

- *Une montée en qualité des données publiques locales (déclaration, normalisation et contenus) ;*
- *Des référentiels de données qualifiés ;*
- *Une aide opérationnelle pour les collectivités (réduction des coûts internes) ;*
- *L'émergence de gisement de données multi-sources fiables.*

Le projet se fixe comme objectifs de développer (moyens):

- *Plusieurs modules de validation, développés soit en spécifique, soit en s'appuyant sur des fonctions externalisées déjà existantes que nous adapterons pour le projet (francisation, adaptation aux standards recommandés, etc.). Ces modules aborderont différents niveaux de normalisation :*
  - *Les contenants : Métamodèles, Formats des fichiers, …*
  - *Les Formats des données (ex : le jeu de données DELIBERATION est-il bien décrit avec les attributs obligatoires que les recommandations ont spécifiés ?)*
  - *Les Syntaxes : structuration des champs (ex : le champ DATE_DELIB est-il bien une date codée au format ISO8601 ?)*
  - *La cohérence de certains champs pivots (ex : ce numéro INSEE_BENEFICIAIRE correspond-il bien à une entreprise/association existante ?)*
- *une plateforme de moissonnage (collecte temporaire des données testées) ;*
- *un ordonnanceur des modules distincts de validation des données référencées ;*
- *Un outil de diagnostic et de certification des différentes sources ;*

- *Une base intermédiaire volatile permettant la consolidation des données à des fins de traitement et d'analyse.*
- *Une interface capable d'alimenter un futur Observatoire de la Donnée locale (quantité, contenus, qualité).*

La plateforme et les outils développés seront mis en œuvre en 16 mois dans le cadre du PIA, avec les partenaires suivants :

- *Portage Administratif : Toulouse Métropole*
- *Portage Opérationnel délégué : Opendata France*
- *Expertise fonctionnelle sur la qualité des données (vision Producteur des données et réutilisateurs « simples ») : FING*
- *Expertise sur les besoins des réutilisateurs (vision Data Scientist/Analyst) : DATACTIVI.ST*
- *Expertise technique pour la conception et l'ingénierie de la plateforme : JAILBREAK*
- *4 Territoires pilotes : Toulouse Métropole, région PACA, Digne-les-Bains, GIP territoire Numérique Bourgogne Franche-Comté*
  (D'autres territoires d'expérimentation pourront être associés au projet dans un second temps).

DESCRIPTION DÉTAILLÉE DU PROJET

1 La solution proposée

1.1 Périmètre, produits, services visés par le projet.

Le projet consiste en la création d'une plateforme nationale de qualification des données publiées en opendata.

La plateforme :

- Effectue des contrôles sur les données publiées par les collectivités (moissonnage et test de jeux de données considérés prioritaires) ;
- Restitue aux collectivités des statuts détaillés sur la qualité des données testées ;
- Évalue et attribue à chaque jeu de données un indicateur de qualité qui permettra de considérer cette source comme suffisamment « propre » pour pouvoir être exploitée dans des traitements de niveau national ou multi-sources (référentiel territorial) :
    - Pour injection dans des bases nationales de référence (ex : data.gouv.fr) ;
    - Pour traitement à des fins de statistiques ou d'évaluation (ex : Observatoire de l'open data territorial) ;
    - Pour analyse thématique ou scientifique (ex : étude comparée sur tels jeux de données : budget, subvention, etc.) ;
    - Pour utilisation combinée avec d'autres jeux de données d'autres producteurs (gisement de données locales dans le cadre de projets de type SmartCity).

Son ambition est de permettre l'évolution qualitative des données disponibles en opendata en apportant une aide opérationnelle aux collectivités et en identifiant les données qui peuvent être raisonnablement utilisées dans des bases consolidées (en effet, l'injection de données de mauvaise qualité peut détériorer significativement des bases déjà constituées à partir d'autres sources).

Cette plateforme ne se donne pas comme objectif de devenir un entrepôt de données locales, ni un référentiel national.

Elle apporte une contribution technique dans l'industrialisation de l'opendata.

Cette plateforme pourra être utilisée pour la qualification d'autres sources (privés, tiers, …), par exemple, l'émergence d'un Service Public Territorial de la Donnée (rapport Smart City de Luc Belot) pourrait en tirer profit pour garantir la cohérence de tels gisements locaux. Ce n'est cependant pas l'objectif immédiat de ce projet d'amorçage.

Tous les logiciels produits seront publiés, dès leur conception et les premières réalisations, en OpenSource.

1.2 Description détaillée de la solution adoptée, innovations et ruptures technologiques impliquées.

L'objectif du projet est de construire, de façon itérative, un prototype opérationnel qui permettra de passer facilement à l'échelle dans une étape ultérieure.

1.2.1 Contribuer à la mise à disposition en open data de jeux de données de référence dans les collectivités territoriales, selon les dispositions de la Loi Lemaire

Les jeux de données du Socle Commun des Données Locales sont ceux ayant un fort potentiel d'impact économique et social, et sont donc le socle indispensable d'une politique d'open data cohérente des collectivités territoriales. S'assurer que ces données sont publiées dans un format standard permet de les rendre comparables et de les agréger au niveau national.

a) La définition des 10 jeux de données prioritaires à mettre à disposition par les collectivités territoriales :

   C'est l'objet d'un travail mené par OpendataFrance dans le cadre du projet OpendataLocale. La définition et la normalisation de ces données est établie et sera publiée au premier semestre 2017. Notons que le Catalogue des Données – qui référence les jeux de données d'un acteur donné – est la première des données normalisées et que ce sera le point d'entrée de la plateforme de qualification. Au moins dans un premier temps, trois autres points d'entrée seront proposés pour faciliter le démarrage de la plateforme : injections manuelle de jeux de données, référencement manuel des jeux, usages des APIs des éditeurs (CKAN, OpenDataSoft, etc.).

   (Voir source ODF : http://opendatalocale.net/index.php/presentation-de-chaque-jeu-et-de-sa-normalisation-precise/)

b) Définition des traitements pour les jeux de données prioritaire :

   Parmi les 10 jeux de données prioritaires, nous sélectionnerons dans le cadre de ce projet d'amorçage 3 à 4 jeux de données sur lesquels la plateforme de qualification mènera un ensemble de tests. Le choix des données et la spécification sera faite par l'équipe de projet avec une perspective métier portée par OpendataFrance, la Fing et Datactivi.st, une perspective technique menée par Jailbreak et la validation par des territoires-pilotes.

Nous choisirons des jeux qui portent sur des problématiques différentes.

1. Un jeu simple, sorte de « hello world » de l'open data : un jeu facile à produire par de nombreux acteurs, facile à spécifier et facile à contrôler ; il s'agirait donc une sorte de jeu de test mais avec de vrais données. Ce jeu pourrait par exemple être la liste annuelle des prénoms des nouveaux-nés : son intérêt peut paraître anecdotique mais c'est une donnée très appréciée du grand public et facile à manipuler et produire. Nous pourrions aussi envisager de faire le choix proche de la Grande-Bretagne qui a choisi comme « hello world » la géolocalisation des toilettes publiques.

2. Un ou deux jeu(x) plus complexe(s) et partagé(s) par plusieurs niveaux d'acteurs, permettant ainsi de comparer les problématiques de qualité et de production selon les acteurs. Il pourrait par exemple s'agir des budgets qui sont produits par tous les niveaux de collectivités.

3. Un ou deux jeu(x) choisi(s) pour leur complexité : par exemple le grand nombre de champs, le volume de données, la piètre qualité des données actuellement produites.

Il faut noter que les tests peuvent porter sur différents niveaux et que leur nature et leur nombre peuvent devenir très importants :

- déclaration : conformité de catalogue au regard du format national spécifié
  - Contenant : le format et la nomination du catalogue est correct
  - Contenu : présence des champs obligatoires, contrôle du format de ces champs, …
  - Cohérence des données qui y sont inscrites : une date est contemporaine, une url est valide, etc.
- Pour chaque jeu de données prioritaire et testé :
  - Contenant : les formats et la nomination du fichier sont corrects
  - Contenu : présence des champs obligatoires, contrôle du format de ces champs, …
  - Cohérence des données qui y sont inscrites : une géolocalisation est bien en France, un numéro de SIREN ou une Adresse correspond à une entrée dans la base nationale SIRENE ou BAN, …

Nous choisirons de commencer par des tests sur les critères les plus importants de notre point de vue. Un travail de concertation sera mené pour en faire le choix.

Un travail collaboratif national sur la qualité des données :

Nous nous baserons notamment sur les travaux menés depuis plusieurs mois par la Fing sur la qualité des données (en partenariat avec l'Ademe, La Poste, la MAIF, la Région PACA et le SGMAP) : http://infolabs.io/sprint-qualite.

Ces travaux ont notamment identifié 117 points de contrôle de la donnée. Les points de contrôles couvrent un spectre très large et portent sur l'enveloppe (le fichier), les métadonnées, la syntaxe, la sémantique, la pertinence, la réglementation, le manque et même la surabondance de données. Ces 117 points de contrôle sont d'ores-et-déjà consultables dans le document suivant :

https://docs.google.com/spreadsheets/d/1XQnBEzdrwVsad51rKuVyyaSHw_6gJwN0JlUjBTOskhc/edit#gid=847338425

À ce jour, une quarantaine de points de contrôle ont été identifiés comme automatisables (par exemple l'encodage d'un fichier ou bien la conformité d'une date au format ISO-8601). D'autres points sont semi-automatisables. La plateforme automatisera au mieux ces contrôles.

La plateforme de qualification des données locales sera l'occasion de préciser et renforcer ce travail. En collaboration avec des acteurs pilotes, nous travaillerons prioritairement sur les indicateurs qualité qui ont le plus d'impact pour les réutilisateurs et qui sont les plus simples à corriger pour les acteurs. En parallèle du développement de la plateforme, nous mèneront un important travail de communication et de sensibilisation aux bonnes pratiques et indicateurs qualité, notamment à travers la communauté Open Data France. Par exemple, la Fing a d'ores-et-déjà produit un document didactique pour faciliter la production d'un fichier CSV de qualité : http://infolabs.io/csv-de-qualite. Nous poursuivrons cet effort avec pour objectif de faire monter progressivement en compétence la communauté des producteurs de données ouvertes.

Ces travaux sur la qualité des données seront publiés indépendamment de la plateforme sous Licence Creative Commons CC-BY.

La plateforme automatise les enchaînements des tests en s'appuyant sur des modules autonomes et des Ordonnanceurs de traitement. Des API seront disponibles pour permettre aux producteurs de données d'utiliser de leur propre chef ces modules autonomes pour tester leurs jeux de données selon leur besoin de validation.

c) Mobilisation du réseau des collectivités territoriales :

- OpendataFrance sollicitera le réseaux des collectivités territoriales pour que les collectivités s'enrôlent dans cette expérimentation : mise à disposition des données prioritaires aux formats normalisés, feedback sur la pertinence des statuts et détails des tests de qualification, amélioration des données publiées pour vérifier la progression des indicateurs et le passage des données dans un état de validité « national ».

- Le moissonnage sera effectué à partir d'une action volontaire des collectivités qui se déclareront sur la plateforme : nom de la collectivité et contact pour les feedbacks de l'outil de qualification, lien vers le catalogue de données qui représente le point d'entrée unique,…

d) Travail avec des collectivités pilotes

Plusieurs collectivités pilotes seront impliquées dans le projet afin de valider la démarche théorique, les éléments concrets de moissonnage, ainsi que les tests des jeux de données référencés. Les collectivités feront évoluer la qualité de leurs données jusqu'à que ces données atteignent un seuil suffisant pour être injectées dans des bases communes. Les retours d'expérience seront prises en compte pour adapter les seuils, les exigences et les traitements.

1.2.2 Construire une infrastructure de validation de la qualité technique des jeux de données open data de référence des collectivités territoriales

Afin de s'assurer que les données mises à disposition sont effectivement exploitables, il est indispensable de s'assurer que les formats sont bien respectés et que les champs pivot (ex: code postal, code INSEE des communes, numéro SIREN, adresse...) sont valides et uniformes. Cette infrastructure permettra de le faire de façon automatique.

a) Conception, développement, mise en place et administration d'une API de validation de fichiers pour les jeux de données contrôlés

b) Conception, développement, mise en place et administration d'une interface web de validation de fichiers pour les jeux de données contrôlés, permettant à toutes les collectivités territoriales de déposer leurs propres fichiers et d'identifier directement les erreurs à traiter

c) Mobilisation du réseau des Collectivités territoriales pour la diffusion et l'utilisation de l'interface web

d) Mobilisation des éditeurs de progiciels et de plateformes de publication d'open data pour la diffusion et l'utilisation de l'API

1.2.3 Construire des indicateurs pour suivre l'état d'avancement de l'open data dans les collectivités territoriales

Il sera nécessaire de suivre l'état de application des dispositions de la Loi Lemaire et de pouvoir baser des recommandations de politiques publiques sur des données fiables et actualisées régulièrement.

Ces indicateurs permettront une vue d'ensemble de l'avancement de l'open data local en France : Conception, calcul et visualisation d'indicateurs de suivi de l'état d'avancement de

l'open data dans les collectivités territoriales (prototype en vue de la mise en place de l'observatoire Opendata Territorial).

1.2.4 Faire avancer l'état de l'art en termes de techniques d'industrialisation de l'open data

L'open data doit passer d'une pratique artisanale à un processus industriel, en raison des obligations juridiques, des masses de données disponibles, de la multiplication des sources et d'un besoin d'actualisation régulière. Pour répondre à ce nouveau défi, il est nécessaire d'utiliser des nouvelles techniques pour mettre en place des flux continus de données.

a) Mise en place d'une chaîne d'intégration continue (moissonnages, validations, agrégation, ordonnanceur, etc) ;

b) Mise en place d'un tableau de bord de visualisation des différentes étapes de l'intégration continue des données ;

c) Documentation des bonnes pratiques et création des conditions pour l'apparition d'une communauté de contributeurs.

1.2.5 Offrir aux usagers une première approche des données ouvertes et validées

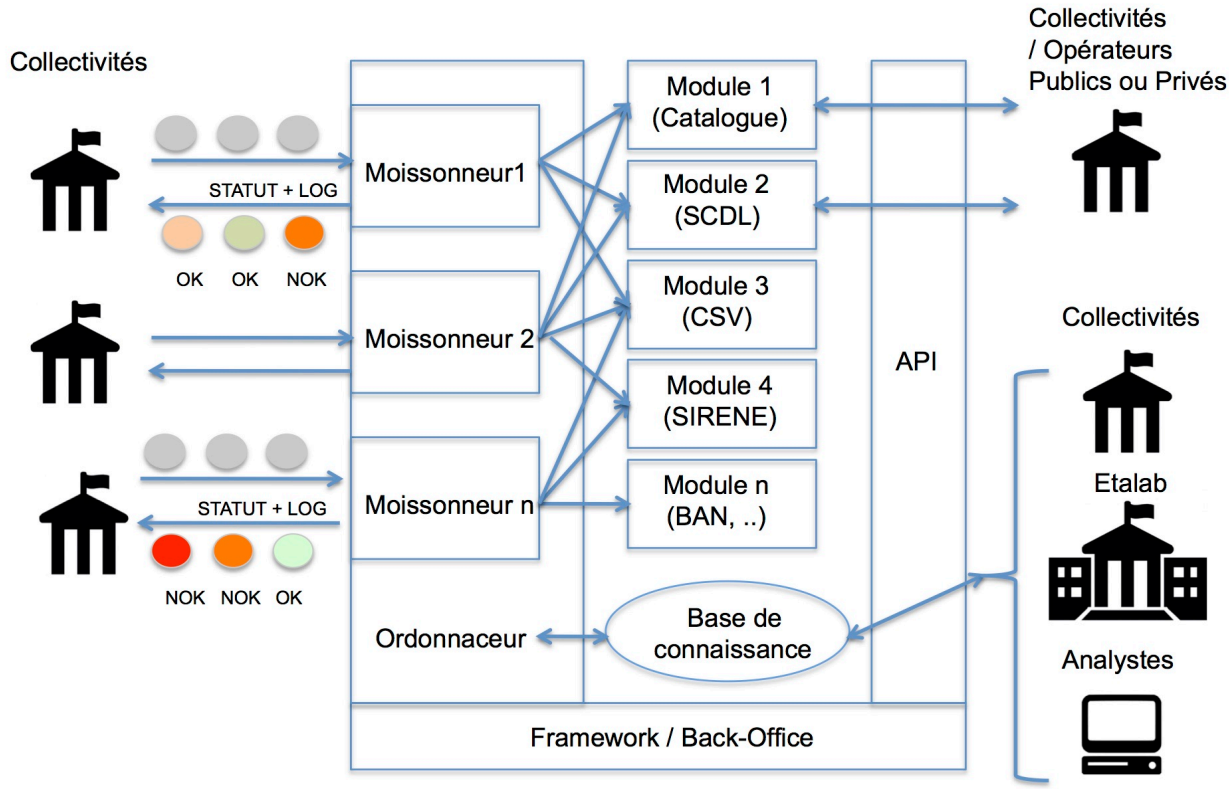
Au-delà des tests de qualité qui peuvent être automatisés, la pertinence de la qualité de données ouvertes s'apprécie par les usages qui peuvent en être concrètement faits. De plus, l'intérêt de constituer des jeux de données consolidés au niveau national (voir au-delà) n'apparaît aux usagers « grand public » que si on leur offre la possibilité d'interagir aisément et rapidement avec ces données.

Aussi, nous proposons de mener, pour chacun des quatre jeux de données qui seront sélectionnés prioritairement dans le cadre de cette expérimentation, un projet d'analyse transverses ou de visualisation de ces données multi sources. Celui-ci permettra à la fois d'avoir, en conditions réelles, des retours sur les spécifications retenues pour les jeux de données et de les améliorer, et de fournir un livrable valorisant l'intérêt de cette spécification, tant pour les producteurs que pour les citoyens.

Le choix des projets de valorisation des données se fera en fonction des thématiques qui seront retenues par l'équipe projet. Le souhait serait de pouvoir mener des projets de complexité croissante, depuis de la simple visualisation interactive d'un jeu de données jusqu'à du machine learning permettant de mettre en évidence le passage de données structurées à de l'information, voire de la connaissance.

Le code source de ces valorisations sera publié en open source.

1.2.6 Architecture fonctionnelle et technique



1.3 Etat des lieux, enjeux et difficultés présentes.

1.3.1 Etat des Lieux

Les données publiées actuellement sont globalement d'un niveau de qualité assez médiocre, surtout du fait de la non-normalisation des données au niveau national. Par ailleurs, les métiers producteurs de données n'ont pas encore intégré que les données qu'ils produisent sont utilisées au-delà de leur métier (silot) et que s'ils savent s'accommoder de la non qualité des données dans l'exercice de leur travail quotidien, il n'en n'est pas de même dès lors que l'utilisation de ces données dépasse leur silot.

Il suffit par exemple de comparer les données de Budgets produites par de grandes métropoles françaises : s'il est incontestable que les données élémentaires sont justes, leur

format de publication rend concrètement impossible un traitement collectif au niveau national sans un lourd traitement de normalisation qu'il vaudrait mieux faire à l'origine.

Cette situation va s'amplifier par l'arrivée des très nombreuses collectivités qui publieront leurs données en application de la loi République Numérique et plus généralement par l'accélération de l'ouverture des données par tous les acteurs publics ou privés. Certains n'auront pas les moyens techniques ou méthodologiques pour produire des données de qualité, cohérentes au niveau national.

Le risque est donc de voir une situation chaotique se développer : Catalogues non normalisés, fichiers (contenants) et données (contenus) non normalisés et mal formatés, données incohérentes.

Pris séparément, cette ouverture porte en elle des vecteurs très positifs pour la transformation de l'action publique. Cependant, il est indispensable de favoriser la réutilisation des données et cela passe par des données normalisées et cohérentes.

Ce travail sera à faire par chaque collectivité et les travaux d'Opendatafrance sur le Socle Commun des Données et sur les outillages d'extraction y contribuent.

Ce qui est en jeu ici, c'est d'identifier les données qui ont atteint un niveau de qualité suffisant pour pouvoir être exploitées simplement au niveau national : application de réutilisation par exemple ou injection dans des référentiels nationaux sans en altérer leur valeur. Cela permettra par exemple d'alimenter data.gouv.fr avec des données qualifiées ou de mettre en place un observatoire de l'opendata territorial avec des indicateurs fiables.

De façon plus générale, les procédures de moissonnage existantes actuellement (niveau national ou sub-national) sont handicapées car elles ne peuvent pas contrôler la qualité des données collectées. Les critères de qualification eux-mêmes ne sont clairement définis.

Il n'existe pas d'outils disponibles pour que les collectivités testent simplement la qualité de leurs propres données. Cette plateforme de qualification donnera aux collectivités –et autres producteurs- des outils externes pour tester la qualité des données qu'elles publient. Il s'agira d'une sorte d'un outil d'audit externe par rapport à des critères nationaux.

Au niveau international, il existe des solutions proches de notre projet, sans la dimension d'automatisation cependant, relativement complexe et de langue anglaise :

http://validator.opendata.esd.org.uk porté par l'ONG Local Governement Association

La construction de notre plateforme s'appuiera sur les implémentations ou les pratiques qui nous paraissent les plus intéressantes, nous assurerons bien entendu une appropriation maximale dans le contexte français : langue, réalité des différents niveaux hiérarchiques

dans les collectivités, concentration sur les données les plus utiles et le socle commun de données locales.

1.3.2 Difficultés présentes

Il n'existe à ce jour pas de solution disponible pour les collectivités sauf quelques outils de validation plus ou moins pérennes et accessibles (par exemple CVSLink CSV Lint – http://csvlint.io/ –, stand-alone et en anglais pour la conformité des fichiers CSV).

Aucune n'est conforme aux recommandations de normalisation Opendata en France et aucune plateforme d'ordonnancement conduit des tests de natures différentes sur un ensemble cohérent de données.

NB : Il sera possible à travers une API ouverte d'activer des tests élémentaires de chaque jeu de données (voir croquis de l'architecture). Les portails de publication, nationaux ou locaux, publics ou privés, par exemple Data.gouv.fr, pourront utiliser cette plateforme pour des tests de leurs propres jeux de données.

Le nombre de données élémentaires à valider est potentiellement très importants ; nous nous concentrerons sur les tests qui possèdent la plus grande valeur ajoutée.

1.3.3 Enjeux

Grâce à cette plateforme, nous avons comme objectifs de :

- identifier des sources de données fiables et qualifiées ;
- connaître les données « interopérables » que l'on peut utiliser à des fins de réutilisation : consolidations, analyses, etc ;
- mettre en place des processus continus de collecte, de qualification et d'agrégation ;
- donner un statut de niveau national sur la qualité des données et des sources locales.

1.4 Attentes des acteurs concernés (usagers, administrations).

a - Du point de vue des usagers Citoyens et Professionnels :

- Disposer de données de qualité acceptables : connaître leur existence et leur niveau de qualité ;
- Pouvoir accéder à un référentiel national de données qualifiées (par ex. où trouver les données publiées en OD sur les budgets ou les délibérations ?) ;

b - Du point des collectivités locales :

- Avoir un outil pour qualifier ses propres données et connaître les erreurs issues de ces contrôles afin de corriger les données et améliorer les processus de gestion et de publication des données.
- Pouvoir nourrir leurs propres outils décisionnels à partir de données de qualité,
- Pouvoir communiquer positivement sur la qualité des données produites par la collectivité,
- Nourrir des bases nationales et obtenir en retour des analyses sur le positionnement de la collectivité.

c - Du point des structures de collecte nationale (Etalab) ou sub-national (portail régional, départemental, thématique ou autre)

- Nourrir le portail (ou la base) de données préalablement qualifiées ;
- Éviter un traitement spécifique pour chaque collecteur ;
- Identifier les producteurs et les sources de qualité ;
- S'appuyer sur un référentiel national aujourd'hui manquant.

d - Du point des vue des acteurs publics ou privés : analystes, acteurs économiques, suivi des politiques publiques,

- Pouvoir constituer des bases nationales homogènes avec des données interopérables et de qualité ;
- Pouvoir faire des traitements multi-sources à des fins d'analyse.

1.5 Conditions et facteurs clés de succès / risques principaux.

Les conditions de réussite et facteurs-clés sont :

- Très bonne connaissance de la réalité des acteurs et des projets Opendata dans les territoires ;
- Sensibilité à la qualité des données ;
- Forte compréhension sur les usages et conditions de réutilisation (dataScientist ou Infolab..) ;
- Expertise technique sur les plateformes de publication existantes (différentes infrastructures de base : ODS, CKAN, …), des outils de test et de qualité, des infrastructures logicielles capables de soutenir ce projet (en opensource) ;

- Connaissance de l'écosystème Opendata en France pour l'animation et susciter la dynamique : participation à l'expérimentation des outils mis en œuvre, contribution à l'amélioration des données publiées, mobilisation des partenaires (producteurs tiers, éditeurs, …) ;

Le facteur-clé réside dans une maîtrise d'oeuvre active qui fédère les partenaires, assure la conduite du projet, rédige les spécifications arrêtées par le groupe projet, mobilise les ressources nécessaire à son développement, gère la recette et la qualité des produits finis.

Les risques principaux sont :

- Complexité des situations : hétérogénéités des données, des formats, des portails, des niveaux hiérarchiques des producteurs

> Mesure de réduction du risque :

- Partenaires reconnus et experts des différents domaines fonctionnels (DataScience, réutilisation, animation territoriale) et technique (moissonnage, normalisation des données, ordonnanceur de traitement, expertise opendata) ;

- Approche incrémentale de la production (agile, livraison régulière, points d'étape réguliers).

- les données à valider sont potentiellement très nombreuses : catalogue, format d'un fichier de données, format des données dans ce fichier, validité (bon format mais mauvais identifiant externe)

> Mesure de réduction du risque :

- dans le cadre de ce projet d'amorçage, le nombre de modules de test est réduit à 4 pour aller jusqu'au bout de l'intégration de chacun. (Si les conditions le permettent, il sera néanmoins possible d'en rajouter) ;

- dispositifs techniques adaptés : il sera mise en œuvre des collecteurs spécifiques pour chaque grande plateforme du marché (OpendataSoft, CKAN, …).

- Architecture trop complexe pour passer à l'échelle

> Mesure de réduction du risque :

- Division du système en modules simples indépendants qui peuvent être autonomes (et des API accessibles directement pour chaque module de qualification) ;

- Faciliter la prise en main et la contribution d'autres acteurs (collectivité, éditeur) pour enrichir les fonctionnalités ou ajouter de nouveaux modules : Mise en logiciel Libre OpenSource au fur et à mesure de leur production ;

- Publication du logiciel en OpenSource : Pérennité des outils et possibilité de prise en main par des tiers (nécessitera un effort soutenu de documentation et d'animation d'une communauté d'usagers et de contributeurs).

1.6 Impact attendu en termes de modernisation de l'action publique.

Les collectivités vivent actuellement un changement profond dans leurs systèmes d'information. Il s'agit de passer de SI « orientés outils » à un SI « orienté données ». La maîtrise des données et de leur qualité permettra cette mutation.

Dès lors que la maîtrise des données sera acquise, il sera possible pour les collectivités de les analyser de façon complète, d'avoir des leviers sur l'analyse de leurs politiques publiques, d'agir sur les facteurs clés afin d'atteindre les résultats souhaités, croiser les données avec l'ensemble des acteurs de leur territoire : autres collectivités, DSP, partenaires privés extérieurs, recherche, université … et évidemment, d'ouvrir les données afin que des réutilisations inédites puissent voir le jour.

2 Méthode retenue

2.1 Présentation des entités porteuses du projet et de la pertinence du partenariat (le cas échéant)

Rappel :

- *Portage Administratif : Toulouse Métropole*
- *Portage Opérationnel délégué : Opendata France*
- *Expertise fonctionnelle sur la qualité des données (vision Producteur) : FING*
- *Expertise fonctionnelle sur le prétraitement et la qualification des gisements des données (DataScientist) : DATACTIVI.ST*
- *Expertise technique pour la conception et l'ingénierie de la plateforme : JAILBREAK*
- *4 Territoires pilotes : Toulouse Métropole, région PACA, Digne-les-Bains, GIP Territoire Numérique Bourgogne-Franche Comté*

La compétence de ces partenaires et leurs contributions sont décrites dans les chapitres concernés tout au long de ce document.

2.2 Principales étapes et méthodologie retenue pour mener le projet (calendrier prévisionnel, jalons et résultats clés…) ;

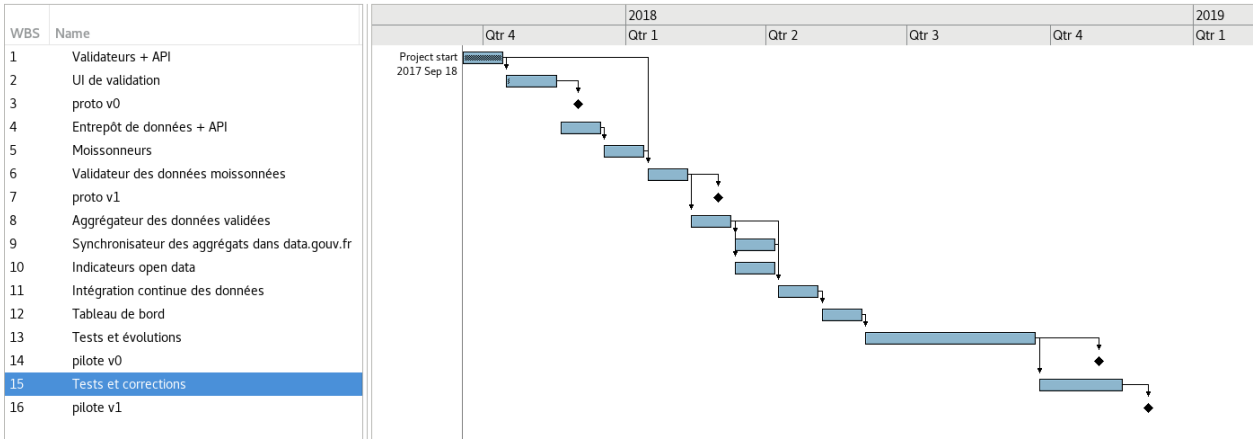
2.2.1 Principaux jalons

Les principaux jalons, établis dans une approche incrémentale, sont les suivants :

- Démarrage du projet : T0 Septembre 2017
- Prototype v0 : T0 + 3 mois Décembre 2017
- Prototype v1 : T0 + 6 mois Mars 2018
- Projet pilote v0 : T0 + 12 mois Septembre 2018
- Projet pilote v1 : T0 + 18 mois Décembre 2018

  fin et bilan du projet d'amorçage.

2.2.2 Planning



2.2.3 Méthodologie

Un projet d'amorçage doit permettre d'investiguer des pistes nouvelles, de tester des approches innovantes, de produire de la connaissance et d'expérimenter des projets pilotes (POC) avec l'objectif d'obtenir rapidement des résultats tangibles pour orienter les phases suivantes de développement et d'industrialisation.

Pour cela, la méthode de développement doit être agile : implication des réutilisateurs dès la conception du projet, prototypage rapide, réutilisation maximale des bonnes pratiques et outillage existants, etc.

Les partenaires du projet sont habitués à de telles méthodes (Scrum, …) et un plan de développement sera établi en début de projet sur ces bases-là.

Le projet sera rythmé par quatre livraisons à intervalle régulier, 2 prototypes et 2 projets pilotes opérationnels.

2.3 Composition de l'équipe (acteurs et compétences mobilisées)

2.3.1 OpendataFrance

Compétences de la structure

> Expertise sur l'opendata dans les territoires ;
>
> Capacité à fédérer les acteurs nationaux (territoires et société civile) et à coordonner les parties prenantes pour l'élaboration des jeux de données cibles et des critères de qualification ;
>
> Maitrise d'ouvrage sur les données nécessaire pour la mise en œuvre de l'observatoire des données locales.

2.3.2 Fing

Compétences de la structure

> La Fing est un think-tank de référence en France sur les usages numériques. Elle travaille depuis plus de 10 ans au développement de l'opendata et en a été un des pionniers avec le projet Réutilisation des Données Publiques Urbaines dès 2008. Elle accompagne depuis de très nombreuses collectivités et acteurs publics dans leur stratégie Opendata.
> La Fing mène de nombreux projets sur la réutilisation des données publiques : Les programmes MesInfos, InfoLab et « parlez-vous data ? » notamment.
> Elle a acquis une expertise sur l'analyse et l'exploitation des données publiques, elle produit des documents de référence sur ce sujet (« comment produire un jeu de données de qualité ») et elle mène auprès de différents acteurs publics ou privés des missions d'animation pour l'amélioration des données ; citons à cet égard sa contribution à la mise en oeuvre du Service Public de la Donnée par Etalab.

2.3.3 JailBreak

Compétences de la structure

> - des stratégies et pratiques d'ouverture par le numérique ;
>
> - des projets menés de bout en bout : depuis la conception jusqu'à la mise en œuvre ;
>
> - du pouvoir d'agir basé sur la capacité à mobiliser les talents là ou ils se trouvent ;
>
> - un réseau à la croisée de différents ecosystèmes: administrations, centres de recherche, entreprises, associations, communautés… ;
>
> - de l'impact international ;
>
> - des méthodes collaboratives et transparentes ;

L'équipe proposée par Jailbreak est constituée de pionniers du libre et de l'open data ayant réalisé le portail data.gouv.fr, le logiciel OpenFisca, l'ouverture du code source du calculateur impôts, la boite à outil du gouvernement ouvert OGPToolbox.org, des dizaines de moissonneurs et de validateurs.

2.3.4 Datactivist

Compétences de la structure

- Expertise sur l'open data, tant d'un point de vue sociologique (processus d'ouverture des données) que du point de vue de la réutilisation des données (analyses quantitatives, machine learning, modélisation, visualisation…) ;

- Bonne connaissance des processus de standardisation de données (Datactivist a travaillé notamment sur l'Open Contracting Data Standard) ;

- Capacité à créer des outils techniques d'exploration, de visualisation ou d'analyse des données pour des utilisateurs grand public ou plus techniques (data scientists, chercheurs).

2.4 Terrains d'expérimentation pressentis.

Les territoires membres d'OpenDataFrance (70 collectivités territoriales) seront sollicités pour participer au projet et co-élaborer les données testées et les critères de qualifications.

Les partenaires territoriaux du projet sont impliqués opérationnellement dans la validation des outils mis à disposition. Pour les collectivités, il s'agit en particulier de :

- *Toulouse Métropole,*
  - *tests de qualification des données produites par la métropole et la mairie de Toulouse ;*
  - *tests de qualification des données produites par les commues limitrophes publiant des données sur la plateforme de la métropole (et par des prodcuteurs tiers publics ou privés dans le cadre du programme SmartCity).*
- *Conseil région Provence-Alpes-Côte d'Azur*
  - *tests de qualification des données produites par la région ;*
  - *tests de qualification des données produites par esa partenaires régionaux publiant des données sur la plateforme OpenPACA ;*
  - *Le région pourra utiliser ultérieurement (c-à-d à l'issue du projet) la plateforme pour les acteurs de la région.*
- *La ville de Digne-les-Bains*
  - *tests de qualification des données produites par la commune.*
- *GIP Territoire Numérique BFC (ex GIP e-Bourgogne)*

- *test des données publiées par les collectivités adhérentes à la structure de mutualisation et apport du cas d'usage sur la publication des données de marchés publics (données essentielles) ;*
- *Le GIP pourra utiliser ultérieurement (c-à-d à l'issue du projet) la plateforme pour ses membres et les acteurs de la région.*

- *D'autres collectivités ont manifesté leur intérêt pour cette plateforme : région Ile-de-France, région Normandie, …*

Elles participent à la conception et à la validation des critères de qualification, aux conditions de déclaration sur le front-office de la plateforme (identifiants et catalogue), aux tests de qualification : les rapports de tests sont-ils pertinents, explicites, cohérents ? etc.

Ces collectivités interagissent avec l'équipe projet pour mettre à disposition des jeux de données progressivement améliorés permettant de vérifier que les outillages aident les collectivités à atteindre un seuil au-delà duquel les données peuvent être agrégées dans une base de test nationale. Des tests d'exploitation sur cette base multi-sources permettent de valider les hypothèses initiales : format des données, validité des critères et des tests de qualification, passage à l'échelle et condition d'industrialisation.

3 Actions prévues pour pérenniser la solution en cas de succès

3.1 Gouvernance

Le projet s'adresse à l'ensemble des collectivités qui ressentent le besoin de valider la qualité des données qu'elles produisent ou publient. Il s'appuie sur des partenaires dont l'expertise et la contribution permettent de concevoir, de réaliser et de tester une plateforme pertinente et efficace. Le projet concourt enfin à la qualité globale des données ouvertes publiées au niveau national et exploitables par de nombreux acteurs publics ou privés.

C'est dire que le projet doit associer de nombreuses parties, et notamment les réutilisateurs finaux de la plateforme.

Sous le portage de Toulouse-Métropole, porteur officiel de l'offre du groupement, le projet sera animé par OpendataFrance en tant que structure légitime au niveau national pour conduire un projet relatif à l'ouverture des données publiques des territoires et, plus particulièrement ici, œuvrer pour la mise à disposition de méthodes et d'outils d'amélioration de leur qualité.

OpendataFrance fédère dès à présent les collectivités engagées dans l'opendata et ce projet correspond exactement à sa mission essentielle : « agir en faveur du développement de l'opendata dans les territoires ». A l'issue de ce projet d'amorçage, le projet se poursuivra d'une façon ou d'une autre : industrialisation des solutions ayant montré leur opérationnalité, amélioration et enrichissement des dispositifs incomplets, reprise de l'approche lorsque les choix faits n'ont pas donné les résultats escomptés. Pour cela, une évaluation très précise sera réalisée à l'issue du projet.

La relation avec les collectivités est alors une condition essentielle : en début de projet pour définir les besoins, en cours de production pour valider les choix et en fin de projet pour mesurer les avantages d'une telle plateforme. Pour cette raison, le projet associe dans le groupement quatre collectivités de nature différente qui participeront à la définition des besoins, à la spécification des critères de validation des données et à la stratégie de test de la plateforme. Il s'agit de :

- la métropole (Toulouse) : test des données de la métropole, de la mairie et des données des communes limitrophes publiées sur le portail métropolitain ;
- une région (PACA): test des données de l'entité régionale et des plus de 40 collectivités qui publient sur ce portail régional ;
- une commune (Digne-les-bains) : test des données locales auprès d'une collectivité de taille moyennes (20 000 habitants) n'ayant pas les moyens de mettre en ouvre une plateforme structurée de test ;
- GIP Territoire Numérique Bourgogne Franche-Comté : test des données publiées par les collectivités gérées par cette structure (plusieurs dizaines de communes et quelques département).

Vis-à-vis des services de l'Etat, intéressés par l'amélioration générale des données publiques issues des territoires, - citons sans exclusive Etalab, l'Agence Numérique, les ministères concernés (Economie Numérique, CGET, DGCL, ..)- OpendataFrance a déjà mis en place de multiples canaux de coordination dans le cadre de ses missions courantes et dans le contexte particulier du projet OpendataLocale. OpendataFrance rend compte de ses travaux lors des comités de suivi avec le SGMAP/Etalab et des comités de pilotage avec l'Instance Nationale Partenariale (INP). Cela sera poursuivi dans la gouvernance du présent projet.

La relation avec les associations d'élus (AMF, ADCF, France-Urbaine, ADF, ARF) se fait dans le cadre des COPIL INP d'une part et de relations directes avec chacune de ces associations d'autre part. Le fait que le Président de l'association OpendataFrance, Bertrand SERP, Vice-Président de Toulouse Métropole en charge de l'innovation et l'économie numérique soit par ailleurs Vice-Président du groupe de travail Numérique au sein de France-Urbaine permet des échanges réguliers et de grande qualité avec les représentants des grandes collectivités, y compris celles non encore engagées dans l'opendata.

Enfin, concernant les acteurs de la société civile, entreprises et associations engagées, OpendataFrance a intégré dans son groupement trois partenaires qui représentent des réseaux d'experts très pertinents dans l'écosystème opendata en France :

- La Fondation internet nouvelle génération (FING) sur la culture de la donnée, sa promotion (InfoLab) et les nombreuses initiatives portées par cette association pour contribuer à la qualité des données publiques (Sprint Qualité dans le cadre de la mise en œuvre du Service Public de la Donnée).

- Datactiv.st : pédagogie de la donnée (Ecole de la Donnée/School of data), exploitation des données et formation (DataScience et logiciel R)

- Jailbreak : jeune entreprise de conseil et de développement informatique, composée d'anciens collaborateurs d'Etalab, cette société est en étroite relation avec le monde universitaire, avec les acteurs de l'OpenSource en France et celui des CivilTech (via OGPToolBox).

La gouvernance aura donc comme modèle :

- le **comité technique** (COTECH) de suivi avec les membres du groupement (partenaires techniques), sous la coordination d'Opendatafrance.

  → Rythme minimum : mensuel. (en sus des réunions techniques journalières ou hebdomadaires nécessaires à l'avancement technique du projet)

- le **comité de pilotage** (COPIL) avec les commanditaires du projet : Etat, Caisse des dépôts,

  → Rythme minimum : trimestriel

- le **comité de coordination** avec les acteurs publics de référence : Etalab, Instance Nationale Partenariale réunissant les associations d'élus des collectivités et les services de l'état concernés (Agence Numérique).

  → Rythme minimum : 4 sur toute la durée du projet, (soit 1 tous les 4 mois approx.)

3.2 Acteurs institutionnels mobilisés pour pérenniser le projet en cas d'expérimentation réussie.

Les acteurs institutionnels mobilisés pour pérenniser le projet seront les mêmes que ceux impliqués dans la conduite et la gouvernance du projet lors de sa réalisation initiale. Ils

seront les mieux informés sur les enjeux, les résultats obtenus, les difficultés à résoudre, les étapes à venir.

Ils sont décrits dans le paragraphe précédent et sont pour rappel :

- **Etalab** : référent au niveau national de l'ouverture des données, prescripteur et ré-utilisateur des données qualifiées pour l'enrichissement du portail data.gouv.fr ou l'analyse qualitative des données produites dans les territoires.

- **SGMAP** : coordinateur au niveau national des politiques publiques numériques. Cela concerne notamment l'articulation avec d'autres programmes du SGMAP comme les observatoires des politiques publiques ou le DCANT.

- **Instance Nationale Partenariale** : groupe de coordination entre l'Etat et les représentants des collectivités territoriales pour la transmission et le recueil des informations à destination des élus locaux et l'articulation avec les projets portés par ces associations. **Ces associations (AMF, AdCF, FU, ADF, ARF) pourraient être sollicitées pour participer au financement de l'industrialisation de la plateforme qui s'adresse prioritairement à leurs membres.**

- **Les services de l'état concernés** : **Agence du Numérique, CGET et DGCL**, qui ont besoin de mettre en place des observatoires sur l'évolution des pratiques numériques dans les territoires et qui peuvent participer à la mise en place d'infrastructure mutualisée de données au niveau national pour couvrir les besoins des territoires.

- **Les investisseurs public ou privés** : dans leur mission d'accompagnement et de soutien à l'économie numérique. La Caisse des dépôts est par exemple très intéressée par cette plateforme qui lui permettrait d'alimenter un observatoire de données territoriales ouvertes. Un dossier est en cours d'élaboration dés à présent (contact Cédric Verpeaux / CDC).

3.3 Documentation prévue du projet en cours de réalisation.

Dans la culture de l'opensource, les logiciels et la documentation seront versés au fur et à mesure de leur élaboration dans un espace de référence, type Github.

Cette publication permet de garantir le partage de la production en temps-réel et la réversibilité du projet en cas de défection d'un partenaire.

Les spécifications recouvrent notamment :

- l'expression des besoins fonctionnels

- la définition de l'architecture technique
- la spécification technique des logiciels produits
- l'élaboration des critères de validation et de qualification (algorithmes)
- la stratégie de test et de validation de la plateforme

4 Résultats attendus

4.1 Gains potentiels estimés en termes de qualité de service ; Impact financier et économique potentiel.

Comme il a été indiqué dans les enjeux essentiels du projet, le manque de qualité et d'interopérabilité des données publiques est un vrai frein à la réutilisation des données publiées.

La plateforme aide les collectivités à produire des données de qualité par ce qui est finalement un service d'audit externe.

L'amélioration des données qui s'en suivra, et leur identification au milieu d'autres données de moindre qualité, permettra aux réutilisateurs, principalement les acteurs économiques, de développer des services ayant :

- une couverture géographique supérieure grâce à la normalisation des données (augmentation de la part de marché et réduction du coût de production),
- une valeur ajoutée supérieure grâce à l'interopérabilité et la fiabilité des données concernées (amélioration de l'offre)

Cette plateforme permettra aussi aux collectivités de réduire le coût de mise en œuvre d'une solution technique de test de la qualité et de l'interopérabilité de leurs données. Il est difficile d'estimer ce gain mais il est probable qu'en cas d'absence d'une telle solution nationale, les tests ne seront tout simplement pas réalisés et la qualité des données publiées restera très insuffisante et compromettra les développements économiques escomptés.

D'un point de vue strictement «citoyens », la faible qualité (et interopérabilité) des données ne favorise pas la transparence de l'action publique puisque la fiabilité des sources peut être remise en question et l'interprétation des données est presque impossible. L'émergence de données qualifiées apportera une plus value démocratique.