etalab gouv.fr

Réponse à la consultation
CoDesign Data.gouv.fr

Nom : Regards Citoyens	Adresse mail : contact@regardscitoyens.org
[ ] □□□□□□□□□□	Téléphone :
[ ] Particulier	Adresse postale
[x] Je souhaite la publication de ma contribution	[x] Je souhaite être invité aux ateliers de travail

Pouvez-vous préciser brièvement les raisons votre intérêt pour l'open data :
(*citoyen, administration, diffuseur de données, réutilisateur, responsable open data...)*

Notre association citoyenne se consacre à la création d'outils numériques pour valoriser les institutions démocratiques en réutilisant de l'information publique. Afin de permettre à chacun de vérifier ou réfuter les informations que nous utilisons et rediffusons, nos outils sont publiés comme logiciels libres et les données qu'ils traitent sont redistribuées en Open Data (gratuité, formats ouverts et licences libres).

Pour réaliser de nouveaux projets démocratiques innovants, nous sommes naturellement demandeurs de données publiques réutilisables : certains projets ne peuvent voir le jour faute de données fiables publiées, d'autres sont ralentis à cause du fastidieux travail d'extraction de données contenues dans des documents publics.

Les questions du débat

1. **Comment améliorer la collecte et la pertinence des données ?**
   (Sourcing collaboratif, accueil de données crowd-sourcées, indexation collaborative, autres pistes ?)

Étalab doit tirer profit des valeurs de transparence et d'ouverture.

Au vu de la taille restreinte de son équipe et du nombre de ses correspondants "données publiques" dans les administrations, il est quasiment impossible pour ces quelques serviteurs de l'État de cartographier de manière exhaustive les données réellement produites par les institutions publiques. En rendant publique au jour le jour la liste des jeux de données déjà identifiés, leur niveau de granularité et leur calendrier de libération, il serait possible de construire un outil public de crowdsourcing recensant les données détenues par l'administration. En permettant à des utilisateurs extérieurs à Étalab de contribuer au recensement, un tel outil pourrait permettre d'identifier les bases de données peu connues, de documenter les points de blocages qui empêchent la libération de certains jeux, d'informer concrètement les citoyens sur les données existantes, d'affiner la description de ces dernières, voire d'identifier des différences entre existant et publié, par exemple en termes de granularité ou d'exhaustivité.

L'amélioration des mécanismes de collecte des données et de la pertinence de ces dernières passe par ailleurs nécessairement par un gros effort de formation et d'accompagnement des fonctionnaires "producteurs". Afin de pouvoir laisser chacun libre de publier sur data.gouv.fr, il convient de faire

connaître les quelques démarches élémentaires à mener en amont de la publication : suppression de la mise-en-page, rassemblement ou séparation des feuilles d'un tableur, simples conversion vers des formats ouverts, ...

La pertinence des données pourrait également être renforcée en capitalisant sur les retours des réutilisateurs. La plateforme data.gouv.fr pourrait par exemple permettre aux réutilisateurs de partager avec les producteurs des versions corrigées ou complétées des données publiées et ainsi améliorer la qualité des données ou enrichir l'écosystème OpenData.

2. **Comment faciliter la réutilisation et l'exploitation des données ?**
   (Quel degré d'interprétation des données par Data.gouv.fr ? Transformation des formats ? Présentation s d'indicateurs, de tableaux de bords ? Datavisualisations ?)

La majorité des données mises à disposition sur data.gouv.fr ne permettent pas une réutilisation aisée : les fichiers sont en majorité fournis dans des formats propriétaires et sous des formes privilégiant la mise en page à la cohérence ou l'homogénéité des données. Dans de très nombreux cas, les réutilisateurs doivent ouvrir les fichiers fournis, observer l'organisation du document pour y découvrir des données cachées dans de très nombreux onglets et passer du temps à homogénéiser ces données en réorganisant les informations éclatées dans plusieurs feuilles, supprimer des lignes entières de titres et des colonnes de mise-en-page, convertir des colonnes simplement colorées en valeurs réelles, ou encore supprimer des lignes de calculs, des cellules agrégées et des formulaires pseudos-dynamiques.

Le recours systématique à des formats ouverts et "machine readable" favoriserait la réutilisation des données en évitant aux réutilisateurs de devoir franchir la barrière d'une longue investigation aujourd'hui nécessaire au traitement de nombreux jeux de données.

De plus, il permettrait aux administrations d'éviter de rendre publiques des informations oubliées car masquées dans des feuilles invisibles, dans les propriétés des formats de document, ou dans des colonnes repliées non-apparentes.

Afin de valoriser les contenus créés par les administrations à partir des données fournies, il pourrait être judicieux de permettre aux producteurs de pointer des exemples de réutilisations produites en interne à partir de ces données, par exemple par des liens vers des images, des PDF ou autres visualisations interactives.

Ces statistiques produites à partir de certains jeux de données publiques par les producteurs sont souvent très intéressantes. Il est important cependant de publier les données sources de ces statistiques et non ces seules statistiques. En effet, les réutilisations les plus intéressantes se basent surtout sur des données dites brutes, c'est-à-dire des données individuelles à la granularité la plus fine, dans le respect bien évidemment du secret statistique. Les agrégations statistiques, régionales, temporelles peuvent ensuite toujours être reproduites par les réutilisateurs sans pour autant bloquer les réutilisations les plus innovantes.

Afin d'offrir une expérience orientée réutlisation, il serait intéressant que data.gouv.fr offre également un outil simple de visualisation des données brutes. Un tel outil supposerait de vérifier quelques éléments de base avant de rendre public un jeu de données (absence d'onglet, absence de cellules fusionnées, conversion automatique en format ouvert réutilisable par des machines). Une telle fonctionnalité pourrait assurer la publication de données de qualité tout en stimulant les producteurs en mesure de visualiser leur travail. Cet outil permettrait si il était mis à disposition des réutilisateurs de leur permettre de mieux apprécier le contenu des données fournies et de voir si elles correspondent à leurs besoins.

Enfin, même en connaissant l'existence d'un jeu de données publié sur data.gouv.fr, il est aujourd'hui extrêmement compliqué de le retrouver au travers de l'ensemble du catalogue. Le moteur de recherche apparaît peu performant en comparaison avec les technologies libres actuellement disponibles. De plus, la démultiplication des éléments présents dans le catalogue rend concrète l'expression "rechercher une aiguille dans une botte de foin". Il est plus que souhaitable de recourir à des technologies comme "solr"[1] ou "elasticsearch"[2], largement utilisées dans la communauté Open Data et ayant démontré leur robustesse.

Un effort important doit être consacré par l'équipe d'Étalab pour documenter et former ses correspondants et les producteurs aux bonnes pratiques favorisant la réutilisation des données. À défaut, le site conservera l'aspect "bric-à-brac" qui le caractérise aujourd'hui. Ces efforts devront être intenses dans un premier temps mais ils devraient être rapidement récompensés : l'information se diffusant dans la communauté des producteurs publics, cela permettra de limiter dans un second temps les interventions des équipes d'Étalab à l'animation de la communauté et à la gestion des réponses aux sollicitations d'expertise.

3. Quelle doit être l'expérience utilisateur sur le site ?

(Expérience de recherche de données ? Accompagnement de débats de société ? Espaces collaboratifs ? Espace personnel ? Portail de la communauté open data ?)

La base d'un site catalogue de données tel que data.gouv.fr repose sur un outil de recherche qui doit posséder un moteur à facettes réellement fonctionnel.

Le second enjeu est la pérennité des ressources publiées. Les urls doivent donc être permanentes pour chacune des séries et jeux de données. Dans le cas de séries de données, il faudrait que la hiérarchie et la version des éléments soient devinables dans les url des données et de leurs fiches afin de faciliter l'agrégation des données et ainsi faire gagner un temps précieux aux réutilisateurs.

Actuellement, le site génère beaucoup de "bruit" dû à la fois à la piètre qualité du moteur de recherche et au problème de la démultiplication des jeux de données au lancement visant à faire du nombre plutôt qu'à favoriser la qualité. À défaut de regrouper l'ensemble de ces données similaires au sein de données agrégées ou de séries consolidées, il est indispensable de "regrouper" les jeux successifs et similaires via une homogénéisation de leurs descriptions et de leur indexation au sein du moteur de recherche pour obtenir des résultats de recherche utilisables et filtrables.

Comme indiqué plus tôt, un important travail d'accompagnement et de guidage des producteurs dans leur renseignement des titres et descriptions doit être engagé par Étalab pour assurer l'homogénéité du catalogue et simplifier la recherche.

Enfin, mettre en avant les données correspondant à des thèmes d'actualité ou à des impulsions ministérielles pourrait permettre à de plus nombreux réutilisateurs de s'intéresser aux informations publiques publiées sur data.gouv.fr. Cette mise en valeur ne doit pas venir de la seule équipe d'Étalab. La communication ministérielle devrait pouvoir être partie prenante dans cet effort.

Abstraction faite du problème régulièrement remonté de certificat invalide sous certains navigateurs lors de l'authentification à l'espace personnel de data.gouv.fr, le dispositif aujourd'hui en place pour assurer l'interaction avec les visiteurs et réutilisateurs semble satisfaisant. Les efforts encore à mener relèvent avant tout de l'animation par des réponses plus systématiquement apportées aux sollicitations des visiteurs plutôt que d'une quelconque évolution technique ou esthétique pour cette partie du site.

1 http://lucene.apache.org/solr/
2 http://www.elasticsearch.org/

4. **Comment favoriser la réutilisation et l'innovation à partir de la plateforme ?**
   (Liste de ressources technologiques ? Annuaire de startups ? Outils pour les développeurs ? Exemples de réutilisations possibles)

L'objectif d'une plateforme publique comme data.gouv.fr n'est pas de faire de la publicité pour des entreprises spécifiques ou de créer des situations biaisant la concurrence en labellisant des acteurs particuliers.

Pour favoriser l'innovation, la plateforme doit mettre en valeur uniquement des technologies qui ne favorisent pas de situation monopolistique. Le recours aux formats ouverts devrait donc être systématique.

La mission Étalab doit être guidée par les valeurs républicaines d'égalité des citoyens et de non discrimination. Si elle décide de mettre en valeur un projet particulier ou une technologie spécifique, elle doit donc s'assurer qu'il réponde à ces valeurs. Elle doit proscrire pour cette raison tout partenariat avec des entités monopolistiques.

S'inscrivant dans le cadre de la modernisation de l'État, la mission Étalab devrait en revanche encourager la publication sur le portail des usages des données réalisés par les propres services de l'État ou les collectivités.

À l'image des sites institutionnels étrangers comme le Federal Register[3], permettre de consulter une information mise à disposition de manière automatique sous plusieurs formats devrait également permettre une réutilisation plus importante.

Enfin, l'intérêt d'une plateforme OpenData, au delà de permettre de simples visualisations d'un jeu de données, est de permettre de recouper différents jeux afin d'en tirer de nouveaux sens au travers de réutilisations innovantes. Proposer des moyens d'identifier facilement à partir de la fiche descriptive d'un jeu de données ceux proposant des champs identiques pourrait par exemple aller en ce sens.

5. **Comment mieux insérer data.gouv.fr dans le réseau des ressources open data ?**
   (Annuaires des ressources nationales ? Liens avec fichiers complémentaires ? Autres pistes ?)

L'un des rôles d'une plateforme nationale est de montrer l'exemple en matière de publication de données publiques à destination de l'ensemble des administrations et collectivités : mener en collaboration avec les différents acteurs un travail de consolidation des formats de données de mêmes natures (budget, recettes, dépenses, marchés publics, transports, compte-rendus d'assemblées, ...) et participer à l'émergence de standards internationaux et à leur adoption par les administrations françaises semble un enjeu essentiel. Rechercher de telles formes d'homogénéisation des structures employées à l'étranger pourrait permettre de catalyser des réutilisations innovantes à l'échelle internationale.

Pour mieux insérer data.gouv.fr dans le réseau international Open Data, une refonte globale de data.gouv.fr en ayant recours comme le data.gov américain, le data.gov.uk anglais ou le datos.gob.br brésilien, à la plateforme générique libre CKAN[4] permettrait l'interconnexion des données mise à disposition sur le catalogue français avec des métamoteurs de recherche comme l'agrégateur européen publicdata.eu.

Contrairement aux risques pris par Étalab en 2010 en citant une marque de logiciel propriétaire dans

3 http://federalregister.gov/
4 http://ckan.org/

son précédent marché public[5], il est tout à fait possible de demander à avoir recours à la plateforme CKAN puisque cette solution est fournie en Logiciel Libre. Comme l'a rappelé le Conseil d'État dans un arrêté du 30 septembre 2011[6], si un acheteur public n'a pas le droit de citer une technologie propriétaire dans un marché public, il peut parfaitement imposer le recours à une technologie libre comme par exemple CKAN.

Enfin, puisque l'API de CKAN[7] est déjà utilisée par différents sites Open Data de collectivités territoriales, son recours pourrait permettre à data.gouv.fr de pointer vers les données de ces autres catalogues.

6. **Comment construire un retour vers les administrations qui partagent leurs données ?** *(Enrichissement des données ? Créations de référentiels de coproduction avec les citoyens ? Autres suggestions ?)*

La libération des données publiques ne signifie pas que les administrations doivent perdre tout contact avec leurs productions : elles peuvent parfaitement attendre des retours de la part des réutilisateurs.

Il existe des modèles contributifs qui permettent de construire juridiquement ce retour attendu des administrations. En ayant recours à des licences imposant un devoir contributif, les administrations peuvent conditionner les réutilisations à la republication par les réutilisateurs des données publiques modifiées. La licence ODBL, équivalente pour les données des licences de logiciels GPL et de contenus CC-BY-SA, est ainsi déjà employée par une douzaine de collectivités comme les villes de Paris, Nantes ou Toulouse[8].

Pour que le recours à ces licences soit valable au regard du droit français, il doit s'accompagner d'un mécanisme dit de double licence permettant aux réutilisateurs qui ne souhaiteraient pas contribuer socialement à l'amélioration des données, de le faire en contribuant économiquement à leur mise à disposition par le paiement d'une licence leur offrant la possibilité d'enfermer les données publiques.

En ayant recours à ce mécanisme, l'administration s'offre la possibilité d'avoir des retours pour tous les usages faits de ses données, soit par une republication par le réutilisateur des données améliorées ou enrichies, soit par un contact permettant une transaction commerciale. Un tel mécanisme pourrait donc également constituer une piste intéressante d'ouverture pour les données publiques encore verrrouillées par des licences payantes.

Une autre possibilité de retour vers les administrations repose sur les actions d'animation de la communauté : en répondant, orientant et synthétisant les dialogues initiés avec les réutilisateurs, les correspondants Étalab et l'équipe de data.gouv.fr peuvent permettre d'identifier les bonnes idées ou les enrichissements de données intéressants pour l'administration. Le recours à des rencontres ouvertes entre administration et réutilisateurs peut également être un moyen efficace pour diffuser innovation et bonnes pratiques au sein de l'administration.

C'est aussi le rôle des producteurs de données d'assurer le suivi de leurs publications sur data.gouv.fr en prenant connaissance des retours effectués sur la plateforme et en répondant de manière transparente aux sollicitations.

5 http://www.klekoon.com/boamp/boamp-appels-offres-conception-realisation-portail-etalab-pour-dila-mission-etalab-1509705.htm
6 http://arianeinternet.conseil-etat.fr/arianeinternet/getdoc.asp?id=192208&fonds=DCE
7 http://docs.ckan.org/en/latest/api.html
8 http://www.opendata-map.org/public/opendata-map.csv

Autres remarques, suggestions...

Le site actuel est confronté à une situation un peu délicate en matière de respect des marchés publics. En ayant imposé le recours à une technologie propriétaire en matière de moteur recherche, Étalab a pris un risque important au vu du III de l'article 6 du code des marchés publics qui indique : « Les spécifications techniques ne peuvent pas faire mention d'un mode ou procédé de fabrication particulier ou d'une provenance ou origine déterminée, ni faire référence à une marque »[9]. En imposant le recours à une technologie de recherche propriétaire, le marché a fermé la porte à de possibles propositions ; nous avons ainsi pu rencontrer des entreprises contraintes d'abandonner leur projet de répondre au marché public.

9 http://www.legifrance.gouv.fr/affichCode.do?idSectionTA=LEGISCTA000006132956&cidTexte=LEGITEXT000005627819