



## 15ème législature

<b>Question N° :</b> <b>17496</b>	<b>De M. Stéphane Peu</b> ( Gauche démocrate et républicaine - Seine-Saint-Denis )	<b>Question écrite</b>
<b>Ministère interrogé</b> > Premier ministre		<b>Ministère attributaire</b> > Premier ministre
<b>Rubrique</b> >État	<b>Tête d'analyse</b> >Pour un traitement ouvert des données collectées lors du « Grand débat »	<b>Analyse</b> > Pour un traitement ouvert des données collectées lors du « Grand débat ».
Question publiée au JO le : <b>05/03/2019</b> Réponse publiée au JO le : <b>09/04/2019</b> page : <b>3236</b>		

### Texte de la question

M. Stéphane Peu alerte M. le Premier ministre sur les conditions de transparence de l'analyse des données collectées lors du « Grand débat national ». À cette heure, ce sont d'ores et déjà plus d'un million de contributions citoyennes qui ont été recueillies. Son Gouvernement a présenté ce débat, non comme un simple sondage d'opinion, mais comme une consultation visant à éclairer les orientations des politiques publiques et à infléchir la décision politique. Dans ces conditions, la fiabilité des données, comme celle de leur traitement, acquiert une importance tout à fait considérable. S'il est assez compréhensible, compte tenu de la masse des données en jeu, d'envisager leur traitement par le moyen d'algorithmes, il conviendra qu'une telle démarche devrait conduire à prendre quelques précautions. En effet, confier l'intelligence collective des Françaises et des Français à l'intelligence artificielle des ordinateurs et des logiciels suppose que l'ensemble des citoyennes et citoyens puisse comprendre et apprécier la nature et la forme des traitements que leurs données auront subis. En d'autres termes, puisqu'en informatique, comme en matière de vie démocratique, « le code fait loi », son expression se doit d'être intelligible et transparente. C'est pourquoi M. le député considère que non seulement l'ensemble des données issues du « Grand débat » doivent être couvertes par une licence du type « open data », mais cette transparence doit également s'appliquer aux méthodologies et logiciels qui assureront le traitement de ces données et dont le code doit être ouvert (« open source »). La fiabilité de l'expertise que produiront ces données dépendra de la faculté concrète (et pas seulement de principe) de contre-expertise à la disposition des citoyennes et citoyens. Il lui demande quelles garanties il compte produire pour que cette indispensable transparence puisse être garantie dans les faits, et que chaque citoyen puisse en vérifier le fonctionnement.

### Texte de la réponse

Les modalités de traitement et d'analyse des données recueillies dans le cadre du Grand Débat National sont guidées par les principes suivants : - Traiter l'exhaustivité des contributions, quelles que soient leurs formes : courriers, courriels, cahiers citoyens, débats locaux, plateforme en ligne ; - Etre le plus transparent possible, en rendant accessibles les contributions ; - La restitution sera multiple, pour refléter la diversité et la richesse des débats et points de vue : il y aura donc plusieurs produits de restitution distincts, en fonction des canaux de contribution. Le traitement de ces données ne repose pas sur l'intelligence artificielle mais sur l'intelligence humaine assistée par la puissance des outils informatiques pour parvenir à une lecture et une analyse très poussée dans des délais brefs. Les données issues de la plateforme du grand débat sont ainsi analysées par Opinion Way, qui s'appuie sur le logiciel Qwam dont l'algorithme procède à des rapprochements lexicologiques, qui permettent

ensuite aux équipes dédiées de relever les propositions les plus récurrentes en tenant compte des liens qui peuvent s'opérer entre elles. Les contributions sous format libre (cahiers citoyens, courriers, autres) ont été transmises à la Bibliothèque nationale de France, qui se charge de numériser leurs contenus. Une fois numérisés, les documents manuscrits font l'objet d'une retranscription par un prestataire. Ces données sont transmises pour analyse au consortium piloté par Roland Berger, associé à Cognito et BlueNove, prestataires spécialisés dans la « civic tech », l'intelligence collective et le traitement de données de masse. Le logiciel de deep learning de Cognito construit un référentiel d'analyse sur la base de la lecture des données, et crée un dictionnaire d'analyse au fur et à mesure qu'il « lit » ces contributions. Les propositions déposées sur la plateforme et les comptes-rendus des réunions locales sont d'ores et déjà accessibles en open data sous licence « Etalab ». Un consortium de chercheurs du Centre National de la Recherche Scientifique (CNRS), de Telecom Paris Tech et de l'Institut National de Recherche en Informatique et en Automatique (INRIA) produira également une analyse. L'agence nationale de la recherche a lancé un appel à manifestation d'intérêt pour l'analyse des données par la recherche, afin d'identifier les forces de recherche susceptibles de se mobiliser et les questions scientifiques originales suscitées par ce jeu de données. Cette licence ouverte libre et gratuite : - apporte la sécurité juridique nécessaire aux producteurs et aux réutilisateurs des données publiques ; - promeut la réutilisation la plus large en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données ; - s'inscrit dans un contexte international en étant compatible avec les standards des licences Open Data développées à l'étranger et notamment celles du Gouvernement britannique (Open Government Licence) ainsi que les autres standards internationaux (ODC-BY, CC-BY 2.0) ; - porte une exigence forte de transparence de la donnée et de qualité des sources en rendant obligatoire la mention de la paternité. Cette licence permet la réutilisation exhaustive des données par tout citoyen. L'ensemble des données ouvertes sont également accessibles via une API (interface de programmation), rendant possible un traitement automatisé et facilitant tout travail de contre-expertise. Les contributions envoyées sous d'autres formats (cahiers citoyens, courriers) comportent parfois des informations de nature personnelle (revenus, situations médicales, circonstances familiales...). Leurs auteurs n'ayant pas explicitement consenti à une publication des contributions, il s'avère en l'état impossible de les porter en ligne et de les mettre à disposition du grand public, conformément au Règlement Général sur la Protection des Données, et à la loi Informatique et Libertés. En application du code du patrimoine (régissant les archives publiques), l'ensemble de ces contributions seront toutefois accessibles par la voie de la consultation des archives départementales et nationales aux personnes justifiant de leur intérêt à en connaître (procédure dite de la consultation en dérogation). La confiance manifestée par les citoyens doit être respectée en protégeant leur vie privée et donc leurs données personnelles. Le Gouvernement expertise donc la possibilité d'anonymiser l'ensemble de ces documents, et prendra les mesures appropriées pour permettre que chacun puisse s'approprier leur contenu sans porter atteinte au respect de la vie privée des participants.