Méthodologie d'une analyse du Grand Débat National

Myriam BÉGEL[1]

Guillaume VIZIER[2]

TITLE

Methodology of an analysis of the Grand Débat National

RÉSUMÉ

Début 2019 a été organisé un débat national, visant à recueillir l'opinion des Français sur quatre grandes thématiques. Un an après, nous revenons sur le Grand Débat National pour partager les points essentiels de la méthode utilisée pour notre analyse des contributions en ligne. Devant les centaines de milliers de réponses, nous avons opté pour une approche semi-automatisée nous permettant d'extraire les opinions les plus répandues. Nous présenterons notamment notre pré-traitement des données : la gestion des réponses multiples, des indications géographiques, ainsi que des réponses textuelles.

Mots-clés **:** *analyse, grand débat, visualisation, code libre, traitement du langage naturel.*

ABSTRACT

Early 2019, a national debate occured in France, aiming at gathering the opinion of the French population on four main topics. One year later, we share the main elements of methodology for our analysis of the online contributions. Facing hundreds of thousands of answers, we chose a semi-automated approach, allowing us to extract the widest spread opinions. In particular, we will present our preprocessing of the data: how we handled duplicate answers, geographical indicators, as well as textual answers.

Keywords: *analysis, grand débat, visualisation, free code, natural language processing.*

1. myriam.grand-debat@begel.fr
2. guillaume.grand-debat@begel.fr

Le Grand Débat National a été organisé à l'initiative du Président Macron pendant le premier trimestre 2019 pour recueillir l'avis des Français sur quatre thématiques : « La transition écologique », « La fiscalité et les dépenses publiques », « La démocratie et la citoyenneté », « L'organisation de l'État et des services publics ». Ce débat est une manière de répondre à la colère exprimée par le mouvement des Gilets jaunes initié à l'automne 2018. Les Français ont pu s'exprimer de différentes manières. Dès le 8 décembre 2018, plus de 16 000 communes ont mis en place des cahiers citoyens. Une plateforme officielle permettait de recueillir l'avis des internautes à partir du 22 janvier 2019. On pouvait répondre à un/des questionnaire(s) à choix multiples ou à des questions ouvertes. Près de deux millions de contributions ont été publiées. Chacun était libre d'organiser des réunions publiques et plus de 10 000 sont répertoriées sur la plateforme officielle avec un peu moins de 10 000 comptes rendus. Une centaine de stands de proximité ont aussi permis de recueillir l'avis des passants. Finalement, 21 conférences citoyennes ont été organisées avec des citoyens tirés au sort ainsi que 4 conférences nationales avec des organisations. Si la représentativité de ce débat est sujette à discussion, c'est un exercice d'écoute de grande ampleur qui a permis à tous les Français qui le souhaitaient de s'exprimer. La période d'expression collective du Grand Débat National s'est achevée le 18 mars 2019. Une première restitution a été présentée le 8 avril 2019 et une mise à jour des synthèses a eu lieu au mois de juin avec l'analyse exhaustive des contributions numérisées. Les données de la plateforme ayant été rendues publiques, chaque citoyen peut consulter les propositions. Nous en avons fait notre propre analyse[3]. Nous l'avons rendue publique le 31 mars sur le site data.gouv.fr. Nos analyses sur les questionnaires longs ont été finalisées la veille de la restitution officielle. Nous avons ajouté les questionnaires courts une semaine après publication des données. Notre code est accessible sous licence libre. Dans un premier temps, nous présenterons les données à disposition. Nous expliquerons ensuite le pré-traitement appliqué aux données en vue de leur analyse, pour laquelle nous approfondirons quelques éléments clefs.

1. Données accessibles

Dans cette partie, nous introduisons les jeux de données relatifs au Grand Débat National. Nous parlerons d'abord des données officielles avant d'évoquer brièvement les sources externes que nous avons utilisées. Nous parlerons enfin d'une plateforme collaborative pour l'analyse des textes.

1.1. Mise en ligne

Sur le site du Grand Débat[4], notamment, le gouvernement s'est engagé à publier toutes les contributions des citoyens : « l'ensemble des contributions au débat [...] seront à terme accessibles à tous. Les restitutions de réunions d'initiatives locales, les réponses aux questionnaires, les cahiers citoyens ouverts dans les mairies, les contributions libres, seront progressivement et régulièrement mis en ligne sous licence libre. » Durant le Grand Débat, la plateforme gérant les questionnaires en ligne a régulièrement mis en ligne les contributions. Dès le 31 janvier, le premier jeu de données était disponible sur la plateforme. On retrouve ces mêmes données sur la plateforme de données ouvertes du gouvernement[5], jusqu'au 2 mars 2019. Passé cette date, les données n'ont été mises à jour que sur la plateforme du Grand Débat[4]. La plateforme a cessé de recueillir des contributions le 18 mars 2019. Les dernières données des questionnaires longs ont été publiées le 21 mars 2019 et les données des questionnaires courts ont été rendues publiques en une fois le 8 avril 2019.

Les méta-informations sur les événements locaux ont été mises en ligne aux mêmes dates que les questionnaires longs. Ces méta-informations servaient à répertorier les événements et comportaient

3. Bégel et Vizier (2019a) : https://myriam.begel.fr/grand-debat
4. Gouvernement français (2019) : https://granddebat.fr
5. Etalab (2019) : https://www.data.gouv.fr/fr/datasets/donnees-ouvertes-du-grand-debat-national/

donc par exemple un titre, une adresse, une date. Les organisateurs de ces réunions pouvaient par la suite publier un compte rendu de leurs événements sur la plateforme. Un peu moins de 10 000 comptes rendus ont été publiés. Malheureusement, ils ne sont pas disponibles au format texte. En effet, le format est laissé libre aux organisateurs, que ce soit au niveau du contenu ou du format informatique. On trouve principalement des images, des fichiers pdf ou Word. Récupérer la donnée brute, le texte, de chacun de ces comptes rendus nécessite donc plusieurs méthodes d'extraction. Nous avons fait le choix de ne pas mener ce travail et de nous concentrer sur les données des questionnaires.

Concernant les contributions citoyennes non électroniques, comme les cahiers de doléances ouverts dans plus de 16 000 communes ou les quelques 27 000 courriers et courriels, elles ont été numérisées et mises au format texte pour permettre leur analyse officielle. Cependant, aucune publication n'a eu lieu un an après, ni n'est prévue, comme le souligne l'article de France 2 (2020).

1.2. Sources externes

Pour contribuer en ligne, un compte est nécessaire. Une contribution peut représenter l'avis d'un citoyen, d'un élu, d'une institution, d'une organisation à but lucratif ou non lucratif. La seule information personnelle disponible sur les contributeurs est leur code postal déclaré. Pour vérifier qu'un code postal est valide et s'en servir dans notre analyse géographique, nous n'avons gardé que les codes présents dans la base de données officielle de La Poste (2017).

Pour les visualisations, nous avons utilisé des tracés des entités géographiques et administratives françaises au format GeoJSON mis à disposition sur Github[6].

1.3. Grande annotation

Si certains contributeurs fournissent des réponses laconiques, d'autres au contraire développent des argumentaires détaillés. On peut en déduire que ceux-ci ont l'espoir que leur participation soit prise en compte et lue par un humain et pas seulement noyée dans la masse et traitée uniquement par ordinateur. Mais même si les textes sont uniques, ils se rassemblent autour de thématiques communes. C'est en se basant sur ces deux principes qu'un groupe de citoyens soutenus par les collectifs Code For France et Data For Good a créé une plateforme d'annotation[7]. Tout le monde peut se créer un compte pour prendre le temps de lire et d'annoter des réponses. N'importe qui peut ainsi apporter sa pierre à l'édifice. Les annotations sont ensuite disponibles à chacun pour analyse. Un texte est annoté par plusieurs personnes et les labels ne sont conservés dans l'export des *annotations convergentes* que si les réponses correspondent. La tâche est énorme pour lire les millions de réponses recensées sur leur site. Au 22 mars 2020, 277 995 réponses ont pu être annotées de manière convergente, c'est-à-dire annotées de la même manière suffisamment de fois (au moins 3 fois) pour considérer l'annotation fiable.

Au vu du faible nombre de réponses annotées (moins de 4% des réponses), nous avons écarté la possibilité d'utiliser cette source de données annnexe pour l'analyse des réponses libres. Nous avons tout de même souhaité évoquer cette initiative car elle aborde la problématique sous un angle différent et pertinent.

6. David (2018) : https://github.com/gregoiredavid/france-geojson
7. GA (2019) : https://grandeannotation.fr

2. Pré-traitement

Dans cette section, nous présentons notre méthode et les principales difficultés rencontrées lors du pré-traitement des données, soit en amont de l'analyse proprement dite. Le code source est disponible sous licence libre[8].

2.1. Gestion des réponses multiples

Une première difficulté provient de la présence de doublons. En effet, il était possible d'enregistrer plusieurs contributions pour un même compte. Par ailleurs, il est également possible pour un même individu de créer plusieurs comptes ; toutefois la détection de ce cas de figure est hors de notre portée.

Prenant le cas du questionnaire intitulé « Fiscalité et dépenses publiques », nous avons un total de 186 711 contributions pour 152 476 comptes. Pour ce questionnaire comprenant 8 questions, 5 185 comptes ont fourni au moins 3 réponses au questionnaire. Nous avons également recensé 1 052 comptes n'ayant fourni de réponse à aucune question du questionnaire.

Pour gérer le cas des doublons, ne conserver que la dernière contribution pour chaque compte ne nous est pas apparu comme judicieux. En effet, les contributions liées à un même compte peuvent différer. Nous avons donc considéré l'ensemble des contributions, sans filtrer pour n'en garder qu'une par compte. Cela peut introduire un biais dans notre analyse, surtout si un même compte a été utilisé par plusieurs individus. Cependant, le nombre maximal de réponses apportées au questionnaire par le biais d'un même compte étant de 115 (cas très exceptionnel), soit 0,076% de l'ensemble des réponses non vides, nous avons considéré le biais suffisamment faible pour ne pas filtrer les contributions d'un même compte. Un compte seul a un faible impact mais nous avons conscience que l'ensemble des doublons a un impact non négligeable. A priori, l'intégralité des propositions a aussi été gardée pour l'analyse officielle.

Une autre complication engendrée par le format des données est apparue lors du traitement des Questionnaires à Choix Multiples (QCM). En effet, les réponses à ces questions sont constituées du texte des réponses sélectionnées, concaténées *dans l'ordre de leur sélection*. Cela peut paraître surprenant, car deux utilisateurs ayant sélectionné les mêmes réponses, mais ayant un ordre de clics différent, auront des réponses différentes dans les données brutes. Cela nous impose un pré-traitement supplémentaire pour retrouver les choix des utilisateurs, indépendamment de leur ordre. De plus, notons que certains QCM offraient la possibilité d'une réponse libre et nécessitaient donc un traitement mixte, à la fois un décompte pour les choix proposés et une synthèse des réponses libres.

2.2. Prétraitement pour l'analyse géographique

En vue d'analyser la répartition géographique des comptes, nous avons cherché à associer un département à chacun. En effet, chaque utilisateur de la plateforme du Grand Débat déclare un code postal, donnée que nous avons voulu exploiter pour analyser les différences géographiques dans les réponses.

Notons tout d'abord que, le code postal dénotant le bureau de poste le plus proche géographiquement, il peut ne pas indiquer le département exact d'une commune si elle est reliée à un bureau de poste dans un département voisin (25 communes en France métropolitaine). L'ambiguïté n'étant pas surmontable au niveau du code postal mais l'impact étant faible, nous avons décidé de tronquer

8. Bégel et Vizier (2019b) : https://gitlab.begel.fr/myriam/grand-debat

les codes postaux *valides* aux deux ou trois premiers chiffres, et de considérer cette valeur comme le département associé au compte (à quelques exceptions près : Corse, Saint-Barthélémy et Saint Martin). À noter que le code INSEE, lui, ne présente pas ce défaut puisqu'il est unique à chaque commune et indique le bon département. Toutefois, ce code n'étant pas connu du grand public, il ne peut pas être demandé aux contributeurs.

La liste des codes postaux valides provient des données de La Poste. Les utilisateurs entrant leur code postal dans un champ libre, certaines entrées sont invalides, du fait de fautes de frappe ou de la présence de contributeurs à l'étranger. Reprenant le cas précédent du questionnaire intitulé « Fiscalité et dépenses publiques », nous avons recensé 2809 codes postaux invalides, soit 1,84% des comptes.

Nous avons écarté les contributions de ces comptes dans le cas de l'analyse géographique uniquement. Elles sont donc bien prises en compte pour l'analyse globale.

2.3. Pré-traitement des réponses libres

Afin de pouvoir analyser les réponses libres des contributeurs, nous avons dû traiter ces réponses en vue d'une analyse de masse. En effet, nous n'avions pas les ressources nécessaires à l'analyse de chaque contribution au cas par cas.

Nous avons donc extrait les thématiques récurrentes pour chaque question par le biais de la fréquence d'apparition des mots. Nous avons commencé par grouper les réponses de tous les contributeurs à une même question et supprimer les mots n'ayant pas de signification propre (articles « le », « la », ..., mots de liaison « de », etc.). Afin de regrouper les différentes utilisations d'un même mot ou de mots d'une même famille, nous avons lemmatisé les mots. Il s'agit de remplacer un mot par un représentant de sa racine, par exemple remplacer les formes conjuguées d'un verbe par son infinitif, ou les pluriels et féminins par leur forme masculine singulière. Pour finir, nous avons compté le nombre d'occurences de chaque mot, groupe de deux mots (bigramme), trigramme, et quadragramme. En effet, l'analyse du nombre d'occurences des mots seuls n'est pas assez précise, car ne tient pas compte du contexte. Ce manque est en partie pallié par l'utilisation des bi-, tri- et quadragrammes. Toutefois, cette analyse peut passer outre le sentiment positif ou négatif de la phrase, comme nous l'avons remarqué pour la première question du questionnaire « Démocratie et citoyenneté », où l'expression « plus confiance » apparaît un grand nombre de fois. Cela ne suffit pas pour savoir s'il s'agit de confiance perdue ou d'un regain de confiance.

3. Analyse

Sans revenir sur nos analyses des réponses à chaque question, disponibles en ligne, nous détaillons dans cette partie quelques points importants comme l'abstention ou les visualisations. Pour finir, nous discuterons des avantages et inconvénients de la méthode utilisée pour le traitement des textes libres.

3.1. Abstention

Comme indiqué précédemment en Section 2.1, des contributions vides ont pu être soumises en ligne. Ainsi, nous avons pris soin pour chaque questionnaire d'indiquer le nombre de réponses totalement vides. De plus, aucune question n'étant obligatoire, nous avons aussi analysé le taux de participation pour chaque question individuellement. Nous avons pris soin également de l'incorporer dans les gra-

phiques des QCM[9]. Le taux de participation n'est pas constant et marque l'intérêt des contributeurs pour un sujet. Le questionnaire long « Démocratie et citoyenneté » comporte trois parties : « Vie institutionnelle et démocratique », « Vie citoyenne » et « Immigration et intégration ». Les questions[10] de ces parties ont respectivement un taux moyen de participation de 74,3%, 59,3% et 64,2%. On notera dans la première partie que le plus faible taux de participation à une question est de 55,9%.

3.2. Visualisation

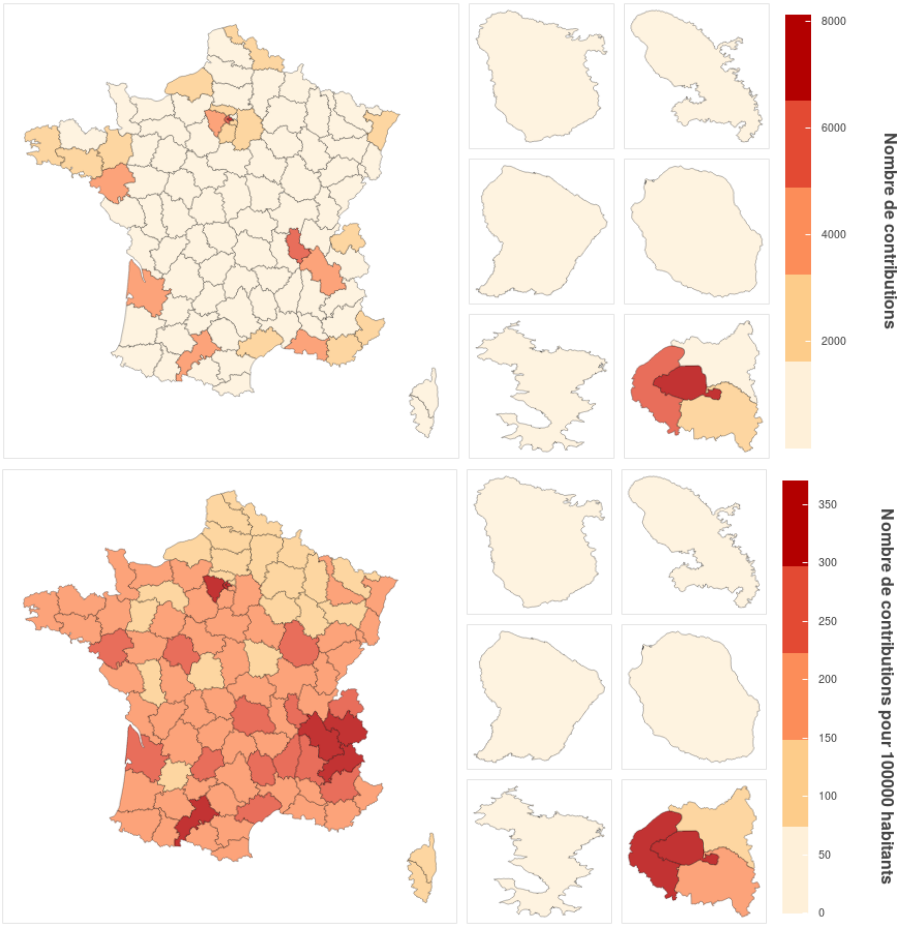

FIGURE 1 – *Cartes par départements des comptes pour le questionnaire long « Transition écologique ». La première indique le nombre de comptes par département. La seconde normalise par le nombre d'habitants.*

Pour chaque questionnaire long, nous avons affiché deux cartes de participation comme en Figure 1. La première est le nombre de participations par département, la seconde normalisant cette donnée par le nombre d'habitants. Les cartes en valeur absolue font ressortir des départements avec de grosses métropoles : Paris, le Rhône, les Yvelines, la Haute-Garonne ou encore les Bouches du Rhône. Les cartes normalisées font ressortir d'autres départements plus ruraux comme l'Est de la région Rhône-Alpes ou le Lot.

Nous avons également regardé la répartition des votes pour les QCM. Certaines cartes montrent seulement que la réponse majoritaire est globalement la même à travers le territoire alors que d'autres montrent des différences intéressantes comme la Figure 2. On y remarque que seul dans

9. Notons par ailleurs que 68,8% des contributions indiquent « Oui » à la question « Faut-il prendre en compte le vote blanc ? ».
10. Hors question de type « Si oui » ou « Si non ».

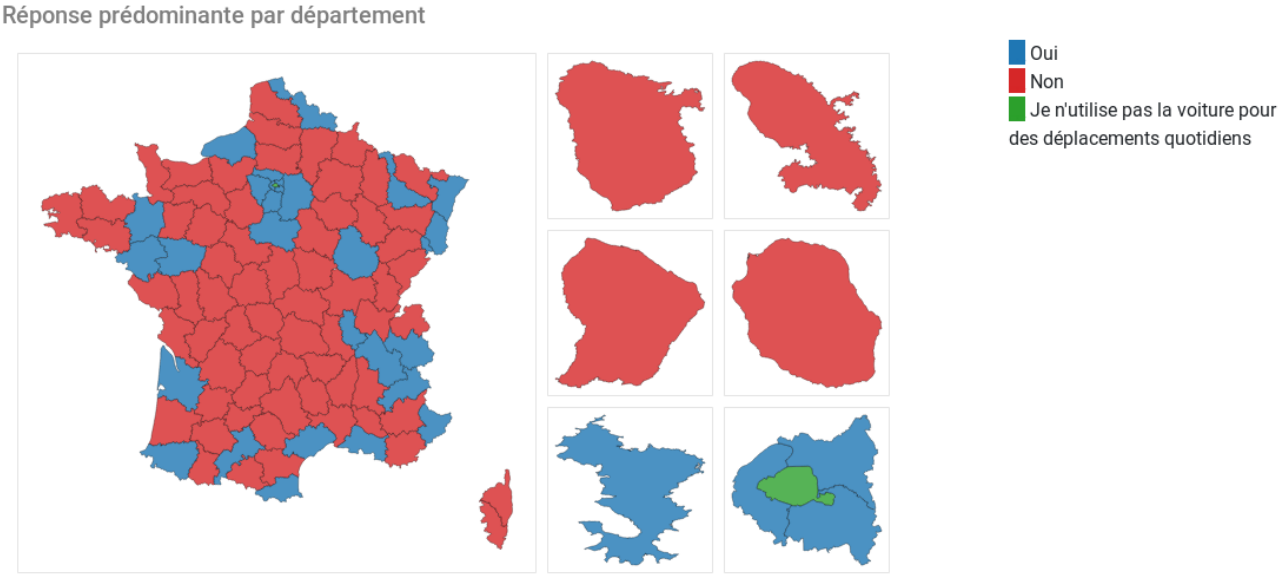

FIGURE 2 – *Carte des réponses prédominantes par département pour la question : Avez-vous pour vos déplacements quotidiens la possibilité de recourir à des solutions de mobilité alternatives à la voiture individuelle comme les transports en commun, le covoiturage, l'auto-partage, le transport à la demande, le vélo, etc. ?*

le département de Paris la majorité des répondants n'utilisent pas leur voiture pour leurs trajets quotidiens. Les départements ruraux ont majoritairement répondu qu'ils n'avaient pas d'alternative à la voiture. Les départements ayant répondu majoritairement qu'ils avaient une alternative à la voiture comprennent une ou plusieurs métropoles.

3.3. Traitement de masse

Notre méthode de synthèse, analyser humainement les n-grammes les plus fréquents dans les textes, présente des avantages et des inconvénients. L'analyse des expressions les plus fréquentes n'est pas aisée car comme nous le disions en section 2.3, il peut manquer du contexte. Par exemple, pour la question « Pensez-vous qu'il serait souhaitable de réduire le nombre d'élus (hors députés et sénateurs) ? Si oui, lesquels », tous les élus sont cités (dans diverses proportions) mais comme nous ne conservons pas le contexte, nous ne pouvons conclure avec certitude si tel élu est plus souvent cité dans ceux qu'il faudrait en effet réduire ou si au contraire il est cité avec une négation.

Les restitutions de nos analyses sont subjectives et non exhaustives. Par ailleurs, nous avons ici traité un effet de groupe. S'il nous est arrivé de lire des contributions pour comprendre le contexte d'utilisation d'une expression, nous n'avons pas analysé des contributions singulières mais nous avons essayé de dégager les opinions les plus fréquentes.

Notre analyse est sensible au lobbying même si nous pouvons parvenir à l'identifier. En effet, souvent les lobbies ont eu recours à des textes identiques. Dans un cas normal, les quadra-grammes sont principalement des expressions qui font sens comme « maire car plus proche » ou « transport commun plus fréquent » et qui peuvent être utilisées dans beaucoup de phrases différentes. Dans le cas d'un lobby, il va y avoir des morceaux plus longs qui se répètent un même nombre de fois. On reconnait ces morceaux plus longs par des n-grammes qui se chevauchent. Par exemple, toujours dans le questionnaire « Démocratie et citoyenneté », il y a environ 565 contributions sur les sourds et malentendants. On retrouve des quadra-grammes comme « langue sourd français faire » ou « sourd français faire choix » ou « faire choix l'inscrire ». Nous arrivons également à identifier des lobbies

par leurs réponses hors contexte comme la limitation des départementales à 80 km/h qui peuvent apparaître dans des questions très éloignées du sujet.

4. Conclusion

Nous avons présenté notre méthode de pré-traitement et d'analyse des réponses textuelles des questionnaires en ligne. Les autres contributions n'ont pas été rendues publiques ou n'ont pas été publiées au format texte. Il subsiste dans notre méthode une partie manuelle pour restituer au mieux le contenu des propositions des contributeurs. Le code Python de notre pré-traitement est disponible en ligne et peut permettre à d'autres d'analyser les n-grammes comme nous l'avons fait.

Pour les comptes rendus officiels de chaque thématique, la méthode utilisée est plus complète et permet de regrouper les propositions similaires et d'en dégager un décompte. Une vérification humaine est aussi effectuée pour assurer la cohérence des groupes de réponses. Les propositions sont gardées si elles sont exprimées par plus de 0,3% des réponses.

Notre analyse, bien que plus succincte, a permis de dégager les opinions majoritaires qui se sont vues confirmer quelques jours plus tard par la présentation des résultats officiels.

Références

(2019), «Grande annotation», URL https://grandeannotation.fr/.

Bégel, M. et G. Vizier (2019a), «Analyse du grand débat national», URL https://myriam.begel.fr/grand-debat.

Bégel, M. et G. Vizier (2019b), «Code de notre analyse du grand débat national», URL https://gitlab.begel.fr/myriam/grand-debat.

David, G. (2018), «France geojson», URL https://github.com/gregoiredavid/france-geojson.

Etalab (2019), «Données ouvertes du grand débat national», URL https://www.data.gouv.fr/fr/datasets/donnees-ouvertes-du-grand-debat-national/.

France 2 (2020), «Grand débat national : un an après, le contenu introuvable des cahiers de doléances», URL https://www.francetvinfo.fr/politique/gouvernement-d-edouard-philippe/grand-debat-national-le-contenu-introuvable-des-cahiers-de-doleances_3784843.html.

Gouvernement français (2019), «Le grand débat national», URL https://granddebat.fr.

La Poste (2017), «Base officielle des codes postaux», URL https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/.