Plan de Gestion des Données, AMULEX

Gökçe Tuncel, Stéphanie Wojcik

PLAN DE GESTION DES DONNEES

ANALYSE MULTIPLATEFORME DE L'EXPRESSION POLITIQUE EN LIGNE (AMULEX)

INFORMATIONS GENERALES

Renseignements administratifs

Acronyme : AMULEX
Code décision : projet de recherche 23R02041S-AMULEX
Titre : Analyse Multiplateforme de l'Expression politique en ligne
Nom de la coordinatrice : Wojcik
Prénom de la coordinatrice : Stéphanie
Affiliation : (Centre d'étude des discours, images, textes, écrits, communication (CEDITEC – EA 3119). / Université Paris-Est Créteil
Contact concernant le PGD : Stéphanie Wojcik et Gökçe Tuncel
Version du PGD : Version n° 1- Juillet 2024

1. DESCRIPTION DES DONNEES ET COLLECTE OU REUTILISATION DE DONNEES EXISTANTES

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les données déjà produites et celles qui seront recueillies sont d'ordre quantitatif et qualitatif et issues de diverses procédures méthodologiques, certaines reposant sur l'utilisation de logiciels spécifiquement dédiées à la collecte de données numériques et utilisées depuis plusieurs années par de nombreuses équipes de recherche.

Quantitatives :

Les données du projet concernent les commentaires en ligne suscités par des interviews de candidat.es à l'élection présidentielle française entre février et avril 2022 sur quatre plateformes (Instagram, Twitch, Twitter, YouTube). Les commentaires portent sur 160 interviews de 12 candidat.es, de durée variable (10 mn à 2h30), disponibles sur des médias privés ou publics, diffusées à la télévision, à la radio ou en ligne. Les commentaires sont issus des espaces en ligne des médias qui ont réalisé l'interview mais aussi des espaces numériques gérés par les équipes de campagne des interviewé.es, où sont diffusés des transcriptions, des extraits et parfois l'intégralité des interviews.
La collecte de 2.062.366 commentaires a été réalisée entre janvier et avril 2022.

Qualitatives : entretiens et questionnaires

Le projet prévoit de constituer plusieurs échantillons d'individus auprès desquels sera administré un questionnaire et qui permettra par la suite de réaliser des entretiens. Le questionnaire sera administré selon deux stratégies complémentaires. Il sera d'abord diffusé de manière ciblée à des internautes sélectionnés dans les espaces de commentaire. Ensuite, il sera mis en ligne sur les comptes de certains éditeurs de contenus (médias et partis politiques) avec lesquels nous aurons conclu un partenariat de recherche. Dans tous les cas, la démarche de recherche, et les finalités des traitements des données personnelles, seront systématiquement explicitées par écrit aux enquêté.es et leur consentement sera recueilli formellement. Enfin, tous les livrables seront rigoureusement anonymisés (modification des pseudonymes et suppression des données sociobiographiques potentiellement identifiantes).

Collecte des données :

La collecte a porté sur des données issues de plateformes qui explicitent la dimension publique de l'expression de leurs utilisateurs et de leurs opinions.

Conformément au RGPD et suivant les recommandations des *Data Protection Officers* des universités et laboratoires participants au projet a été publiée sur leur site une information de principe sur les objectifs de la recherche, à laquelle seront systématiquement renvoyés les internautes avec lesquels nous prendrons contact :

Information affichée sur le site du CEDITEC, Université Paris Est Créteil: https://ceditec.u-pec.fr/presentation/projets-de-

recherche/information-collecte-des-donnees-projet-amulex
Information affichée sur le site de Paragraphe, Université Paris 8 Vincennes Saint-Denis : https://paragraphe.univ-paris8.fr/Projet-ANR-AMULEX

Pour collecter les données elles-mêmes, ont été utilisés des logiciels libres développés par des centres de recherche ou des développeurs indépendants.

Concernant les collectes sur Twitter, elles ont été réalisées à l'aide de *Gazouilloire* conçu et maintenu par le Médialab de Sciences Po Paris (https://medialab.sciencespo.fr/outils/gazouilloire/)
Pour YouTube, a été utilisé un module intitulé YouTube DataTools disponible sur le site du Digital Methods Initiative de l'Université d'Amsterdam (https://ytdt.digitalmethods.net/).

Les commentaires déposés sur Instagram et sur Twitch ont été recueillis grâce à deux logiciels disponibles sur la plateforme GitHub, soit pour les premiers, Instaloader (https://github.com/instaloader/instaloader) et TwitchDownloader (https://github.com/lay295/TwitchDownloader) pour les seconds.

L'ensemble des collectes a été intégré dans une base de données (BDD) constituée par les ingénieurs de recherche recrutés pour le projet. Elle a été déposée sur un serveur fourni par l'un des partenaires du projet, l'Institut des Systèmes Complexes en Ile-de-France (ISC-PIF). Une fois les données personnelles anonymisées, la base de données sera disponible sur l'entrepôt NAKALA géré par le consortium de recherche Huma-Num.

Comment les données préexistantes vont être utilisées :

Pour la collecte de nouvelles données, ce sont les mêmes logiciels qui sont utilisés. Certaines collectes, sur de faibles volumes, sur Instagram ont été réalisées manuellement.

Durant le projet, la base de données n'est accessible qu'aux membres du projet. Une fois le projet achevé, la base de données, anonymisée et déposée sur NAKALA, ne sera accessible que sur demande (une note d'intention d'une page est demandée), à l'aide d'un code d'accès propre à chaque utilisateur, pour une durée limitée qui reste à définir.

Pour les données numériques issues des plateformes, leur provenance sera documentée dans un tableau en format CSV, précisant la source et la date de la vidéo qui suscite les commentaires (émission ou compte professionnel des candidats), la plateforme où ils ont été énoncés, la date et l'heure de la collecte ainsi que l'outil de collecte utilisé.

Les réponses au questionnaire seront insérées dans un tableau Excel et dans un logiciel de traitement de type Statistical Package for Social Sciences (SPSS). Par ailleurs, le tableau comportera le nom de l'administrateur du questionnaire, les usagers contactés pour l'envoi du questionnaire et la demande d'entretien.

Les entretiens seront déposés dans un dossier uniquement accessible aux membres du projet, sur la plateforme ShareDocs gérée par Huma-Num. Un tableau associé précisera les métadonnées telles que la date, l'heure, l'endroit et la personne qui réalisera l'entretien.

Les données collectées par les organisations publiques telles que par exemple la Bibliothèque Nationale de France et l'Inathèque ne sont pas suffisamment nombreuses pour répondre à nos questions de recherche spécifiques. Les données collectées par des entreprises privées, outre qu'elles ne répondent pas à nos questions de recherche, sont uniquement accessibles de manière payante.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Le corpus de données numériques issues des plateformes se compose de 2.200.000 contenus textuels et des fichiers multimédias (audios, vidéos et images) faisant environ 130 giga octets.

Les données produites et recueillies sont de type suivant :

- commentaires extraits des quatre plateformes en tableurs CSV, XLSX, TAB et JSON;
- textes, images et vidéos publiés sur les plateformes (jpg, png, csv, mp3, mp4);
- toutes les données (commentaires extraits, textes, images, son, vidéos) sont stockées dans la BDD en format PostgreSQL;

- réponses au questionnaire collectées et sauvegardées via LimeSurvey, adressées par mail ou par la plateforme du réseau social (Twitter, Instagram et Twitch)
- entretiens réalisés en présentiel ou en visioconférence (comme Teams) enregistrés en format audio et/ou vidéo

Les données textuelles, hors éléments identificatoires des internautes, font l'objet d'un traitement avec les logiciels Prospero et Gargantext.

Les données audios et vidéos seront traitées avec des logiciels de retranscription automatisée fournis par Huma-Num (Whisper) et/ou installés sur les ordinateurs des ingénieurs de recherche et du post-doc du projet, dans le plein respect du RGPD.

Le logiciel Gargantext, produit par l'ISC-PIF (partenaire du projet), et Whisper, produit par OpenAI, sont des logiciels ouverts. Prospéro est un *freeware*, c'est-à-dire que son utilisation est gratuite à des fins scientifiques. Par ailleurs, une partie du budget AMULEX est dédiée à sa reconstruction technique en logiciel open source.

Les données textuelles (.csv et .json) collectées sont intégrées dans une base de données relationnelle de type SQL (PostGreSQL). Les fichiers audios, vidéos et les images seront stockés dans leur format d'origine (mp3, mp4, avi, jpg, png, etc.) dans l'espace ShareDocs dédié au projet, son hébergement est fourni par Huma-Num, auquel seuls les membres du projet ont accès.

Le choix d'une base de données relationnelle répond à la nécessité de rendre ces données interopérables, et de faciliter les comparaisons entre les données issues de quatre plateformes numériques. L'interopérabilité est nécessaire pour l'analyse qui repose sur les différents outils de traitement retenus (Prospero, Gargantext) dans le projet AMULEX.

Dans la seconde phase du projet, nous menons une enquête par questionnaire et entretien. Pour les questionnaires auto-administrés, avec un taux de réponse autour de 3 % des personnes contactées (cf. projet pour le mode d'échantillonnage), nous espérons obtenir 1000 à 2000 répondants pour Twitter, 300 pour YouTube. Pour Instagram, où le volume d'utilisateurs est moins élevé, et Twitch, où l'on ne peut contacter que les *streamers (*éditeurs des vidéos en flux*)*, nous sollicitons les éditeurs de contenus afin qu'ils mettent en ligne le questionnaire sur leur compte ou leur page, dans le but d'obtenir 150 à 300 réponses d'utilisateurs de Twitch et d'Instagram.

En ce qui concerne les entretiens semi-directifs, nous souhaitons en réaliser avec une centaine d'usagers aux pratiques variées, sur l'ensemble du territoire métropolitain, à distance ou en présentiel.

2. DOCUMENTATION ET QUALITE DES DONNEES

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Étant donné que nous travaillons avec une base de données relationnelle, le fichier "lisez moi" des données de la BDD consiste en la description de chaque champ des tables qui compose cette base de données. Ces descriptions seront attachées à la base de données elle-même et seront disponibles dans un fichier texte. La BDD est construite en fonction des principes FAIR : les données doivent être faciles à trouver, accessibles, interopérables et réutilisables. Les principes FAIR constituent la préoccupation collective du projet dans une perspective de science ouverte afin d'assurer la réutilisation des données par la communauté scientifique.

L'entrepôt de la base de données NAKALA respecte également les principes FAIR (pour plus d'information sur NAKALA : https://documentation.huma-num.fr/nakala/)

Une documentation en format texte sera mise à disposition sur HAL en accès libre. Cette documentation comprend une brève méthodologie dans une perspective de réutilisation possible des données. Ce document contient les étapes de la collecte des données sur quatre plateformes (les outils utilisés, les objectifs, les obstacles rencontrés), de la conception des questionnaires et des entretiens (les choix théoriques et méthodologiques, les hypothèses principales, l'articulation avec la problématique du projet, la procédure et la méthode conçues pour le recrutement des enquêtés, les obstacles rencontrés).

Les métadonnées nativement numériques sont les suivantes :

<u>Utilisateur (internaute qui laisse un commentaire en direct ou en différé) :</u>

YouTube : Identité YouTube; pseudonyme affiché sur YouTube, le lien de la chaîne YouTube de l'utilisateur; nombre d'abonné-es et d'abonnements.

Twitter : pseudonyme (@) : photo de profil, localisation du compte, biographie du compte, nombre de tweet que la personne a publié depuis l'ouverture de son compte, nombre d'abonnés et d'abonnements, nombre de likes, statut du compte privé, public ou professionnel.

Instagram : Identité Instagram, pseudonyme, photo de profil, biographie du compte, nombre d'abonnés et d'abonnements; statut du compte privé, public ou professionnel.

Twitch : Pseudonyme; photo de profil, biographie du compte, nombre d'abonné-es et d'abonnements.

Les publications d'Instagram, YouTube et Twitch :

Type de plateforme, type de publication (vidéo, image, texte), langue, date de publication, date de modification, date de maintien en ligne (duration time), nom de chaîne, "caption" (légende de la photo ou de la vidéo. Le texte qui accompagne une publication), nombre de vues, nombre de "j'aime", nombre de "je n'aime pas", nombre de republications, nombre de réponses, nombre de commentaires, statut de "privacy".

Les commentaires :

Type de plateforme, contenu, date de publication, nombre de "j'aime", nombre de réponses.

Les questionnaires et les entretiens :

Les métadonnées des questionnaires et des entretiens concernent : la date de l'entretien, sa durée, son lieu, sa forme (présentielle et/ou distancielle), la plateforme à partir de laquelle l'interviewé est recruté pour l'entretien, administrateur ou enquêteur.

Protocole de collecte de données pour les questionnaires et les entretiens :

La méthodologie de collecte et le mode d'organisation des données seront formalisés et consignés dans les documents suivants :

Tableau des codes enquêteurs pour les entretiens et des codes administrateurs pour les questionnaires

Tableau des codes entretiens et questionnaires

Fiche de codage des entretiens et des questionnaires

Trame d'entretien et de questionnaire

Entretiens et questionnaires – transcriptions et fiches : informations consignées dans le fichier « lisez-moi »

▪ Nommage des fichiers de retranscription d'entretiens et des questionnaires

Code de l'interviewé + date_entretien

Intervieweur

Code de questionnaire + date_envoi + date_réception

Administrateur

Cartouche à renseigner en haut de chaque traitement d'entretiens (retranscription) et de questionnaires :

Cartouche entretien :

Intervieweur : trigramme de la personne ayant réalisé l'entretien déterminé dans le fichier « lisez moi »

Interviewé : déterminé dans le fichier « lisez moi »

Date et heure de l'entretien

Durée de l'entretien

Contexte : entretien téléphonique ou visioconférence ou face à face/

Transcripteur : nom

Relecteur : nom de la personne qui "corrige la transcription"

Date de relecture : date à laquelle la transcription est considérée par le chercheur comme pouvant être exploitée car relue et corrigée

Cartouche questionnaire :

Administrateur : trigramme de la personne qui a administré le questionnaire déterminé dans le fichier « lisez moi »

Répondant : déterminé dans le fichier « lisez moi »

Date et heure de l'envoi et de la réception du questionnaire

Plateforme par laquelle le questionnaire a été administré et reçu (réseaux sociaux ou par e-mail).

Les standards de métadonnées sont ceux de la DCMI - Dublin Core Metadata Initiative

Comme indiqué plus haut, les données seront gérées à travers une base SQL et classées avec les métadonnées indiquées. Il y aura des versions dérivées produites à partir des données brutes traçables. Il s'agira, par exemple, des versions avec une transformation des émoticons en texte : les commentaires seront concaténés selon des critères variés (tels que, par exemple, le fait de provenir d'une même plateforme).

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Les données (brutes) sont recueillies avec des outils dédiés dans les formats CSV et JSON. Elles sont ensuite importées dans la BDD relationnelle. Dans un troisième temps, ces données sont nettoyées afin de supprimer des doublons, ou bien, par exemple sur Twitter, des énoncés qui ne comportent pas le *hashtag* défini au préalable pour la sélection des tweets.

S'agissant des questionnaires, et des guides d'entretiens, ils seront élaborés de manière collective par l'équipe, sous la supervision des responsables de l'Axe 2 dédié à l'enquête sociographique.

3. STOCKAGE ET SAUVEGARDE PENDANT LE PROCESSUS DE RECHERCHE

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

La base de données et son système de stockage seront déposés sur un serveur avec accès physique sécurisé à l'ISC-PIF et sera accessible à distance par une connexion chiffrée SSH.

D'autres données (telles que les données brutes en format CSV, les vidéos des interviews ainsi que les images qui accompagnent certains commentaires, les données provenant des questionnaires et les entretiens) seront sauvegardées sur la plateforme ShareDocs, Huma-Num.

Dans les deux cas, des systèmes de sauvegarde automatiques mensuels seront assurés avec une copie maintenue dans le laboratoire Paragraphe à l'Université Paris 8.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

En cas d'incident, les données de sauvegarde sont exportées dans les emplacements d'origine : ShareDocs Huma-Num et BDD.

Les membres du projet ont accès à toutes les données sur Sharedocs via un compte personnel unique avec un mot de passe. Pour la BDD et l'entrepôt de la base de données sur NAKALA, les accès sont sécurisés et sont soumis à une journalisation : chaque entrée dans la base sera enregistrée dans un journal qui indique les noms, la date et l'heure d'accès.

Les principaux risques concernent l'accès indu à la BDD par des tierces personnes non membres du projet, des entreprises et/ou des acteurs politiques. Ces risques sont maîtrisés, d'une part, grâce au contrôle des accès au serveur (connexion chiffrée SSH) et, d'autre part, par l'anonymisation des données lorsque seront publiées les analyses de l'équipe, et versées dans l'entrepôt de base de données NAKALA.

La responsable scientifique du projet, ainsi que les co-responsables d'axes, représentant leurs institutions respectives ont renseigné une déclaration de traitement déposée auprès des délégués à la protection des données de celles-ci, conformément aux indications qui leur ont été fournies : le laboratoire CEDITEC, l'Université Paris-Est Créteil, le laboratoire Paragraphe, l'Université Paris 8 et l'ISC-PIF.

4. EXIGENCES LEGALES ET ETHIQUES, CODES DE CONDUITE

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Le projet AMULEX respecte ainsi le règlement général sur la protection des données (RGPD) de l'Union Européenne. L'engagement des membres du projet sur le volet RGPD a par ailleurs été stipulé dans la fiche registre transmise au délégué à la protection des données personnelles (DPO) de l'UPEC.

Annonce sur le site du CEDITEC, Université Paris Est Créteil: https://ceditec.u-pec.fr/presentation/projets-de-recherche/information-collecte-des-donnees-projet-amulex
Annonce sur le site du Paragraphe, Université Paris 8 Vincennes Saint-Denis : https://paragraphe.univ-paris8.fr/Projet-ANR-AMULEX

L'information qui serait spécifiquement fournie à chaque internaute s'exprimant sur une plateforme numérique requiert "des efforts disproportionnés". Par conséquent, une information générale est publiée sur les sites web de CEDITEC (UPEC), Paragraphe (Paris 8) et l'ISC-PIF afin d'informer les individus de l'existence du projet, de ses objectifs et des collectes qui ont été réalisées. Dans l'hypothèse où ils s'opposeraient au traitement de leurs données, les internautes peuvent contacter la responsable scientifique du projet afin que leurs commentaires soient exclus de la BDD. Concernant l'information et l'obtention du consentement éclairé des enquêtés, l'annonce d'information disponible sur les sites internet des partenaires présente l'objectif principal du projet ainsi que le protocole de recueil, de traitement et de conservation des données. Cette annonce affiche également le contact des responsables des traitements dans chaque établissement. Toutes les personnes interrogées pour les entretiens par questionnaires et pour les entretiens semi-directifs ont donné leur consentement à l'utilisation de leur propos pour l'enquête. Ce consentement est déclaré oralement en début d'enregistrement pour les entretiens. La lettre de consentement émanant de la CNIL leur a été transmise. Ainsi, conformément à l'article 13 du RGPD, les informations suivantes ont été notifiées aux personnes concernées par le traitement de leurs données :

- finalité du traitement ;
- nom et coordonnées du responsable du traitement ;
- nom et coordonnées du DPO ;
- durées de conservation de leurs données ;
- données concernées et leur utilisation ;
- modalités d'exercice de leurs droits : droits d'accès, de portabilité, de rectification, à l'effacement, de limitation et d'opposition;
- Les enquêtés pourront contacter les membres du projet via l'adresse mail indiquée par le chercheur;

- Le respect des données à caractère personnel et l'anonymat des interlocuteurs sera mis en œuvre dans le cadre de tout processus de publication et de dépôt de la base de données dans l'entrepôt (NAKALA), dans le respect des codes déontologiques en vigueur en Europe.

- En ce qui concerne les réponses d'enquête (transcriptions et enregistrements sonores des entretiens), un code est attribué aux personnes interrogées. Toute référence à des lieux, à des personnes ou à des fonctions qui permettraient de reconnaître la personne est codée.

Les membres du projet, en conformité avec les principes de l'éthique de la recherche, respectent la confidentialité des données sur le terrain (ne pas divulguer les informations personnelles dans le cercle d'inter-connaissances), et durant le processus de recherche (respect de la confidentialité des données).

Les personnes contactées pour les entretiens et les questionnaires ne sont pas dans l'obligation de partager leurs identités personnelles. Les personnes peuvent également exprimer leur souhait de ne pas être enregistrées lors des entretiens. Le fichier audio sera supprimé une fois que la retranscription des entretiens aura été effectuée.

Afin de systématiser la pseudonymisation des individus qui se sont exprimés sur les quatre plateformes et dont nous avons recueilli les commentaires, nous allons d'abord faire une table de correspondance. En revanche, toute forme d'identification, pseudonyme compris, sera supprimée dans un second temps, lorsque la BDD sera mise en dépôt sur NAKALA et dans les publications scientifiques qui découleront du projet.

Seuls les membres du projet ont accès aux données sur ShareDocs Huma-Num et à la BDD. Le droit de modifier et de supprimer le contenu du ShareDocs appartient à la responsable scientifique du projet Stéphanie Wojcik.

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

S'agissant d'un projet coopératif, les laboratoires scientifiques sont responsables des données qu'ils ont recueillies. Ce sont les interlocuteurs dédiés, services juridiques, au sein de ces institutions scientifiques qui appliquent la législation conforme.

Les données sont accessibles aux membres du projet. Elles pourront être accessibles, une fois anonymisées, à d'autres chercheurs, sur demande motivée et pour un temps limité.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

Les enjeux éthiques sont intégrés à la réflexion méthodologique déroulée à toutes les étapes de la recherche (recueil, traitement, stockage, publication des données). Ces démarches ont été coordonnées par les responsables des différents axes du projet.

5. PARTAGE DES DONNEES ET CONSERVATION A LONG TERME

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

À la fin des 36 mois dévolus au projet, les données anonymisées par nos soins seront mises à disposition sur l'entrepôt dédié, NAKALA. La préservation des données sera déléguée à l'entrepôt choisi et la recommandation sur la durée d'archivage sera de 10 ans.

Les entreprises privées, les organisations publiques n'ayant pas pour objet la recherche scientifique et les acteurs politiques (partis politiques, etc.) n'auront pas d'accès à la BDD sur NAKALA.

Les chercheur-es y ont accès, sous condition. La demande est faite via le formulaire de contact disponible sur NAKALA (fonctionnalité "demander l'accès à un fichier sous embargo"). Le gestionnaire des données (responsable scientifique du projet) traite la demande, la valide ou non, et définit ensuite les modalités d'accès aux fichiers. Les conditions d'accès aux données sont les suivantes : la rédaction d'une page qui présente clairement les objectifs de la recherche et l'attachement institutionnel du demandeur. Chaque utilisateur s'engage à mentionner dans les productions scientifiques de tout type la mention suivante : « réalisé avec la BDD AMULEX ».

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

La base de données sera conservée dans l'entrepôt NAKALA après l'anonymisation. Les données et métadonnées qui permettront d'identifier les individus ne seront pas conservées.

Comme indiqué dans le guide NAKALA (https://sharedocs.huma-num.fr/wl/?id=pw007Gnmh206KI9Wggft5dWYGrA497rX), il s'agit d'un entrepôt de données de recherche pour les sciences humaines et sociales. Son accès et son utilisation sont décrits dans la documentation d'Huma-Num : https://documentation.huma-num.fr/nakala/

Dans NAKALA, la description est basée sur un ensemble minimal de cinq informations (Type de dépôt, titre, auteurs, date de création, licence) qui peuvent être enrichies de manière étendue et cumulative. Aux cinq champs obligatoires de la notice de description de NAKALA (propriétés nakala), il est donc possible d'ajouter et dupliquer tout autre champ issu du vocabulaire Dublin Core qualifié qui est à disposition dans les zones "Informations complémentaires" et "Ajouter d'autres métadonnées"

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Les données étant partagées sur l'entrepôt NAKALA, l'accès pourra se faire par un navigateur. Nous mettrons en place une procédure d'identification et demande d'accès pour consulter les données sur NAKALA.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

L'entrepôt NAKALA attribue des identifiants pérennes à chaque base de données.

6. RESPONSABILITES ET RESSOURCES EN MATIERE DE GESTION DES DONNEES

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

La responsable scientifique du projet Stéphanie Wojcik (CEDITEC, UPEC) est la responsable légale des données. Orélie Desfriches Doria (Paragraphe, Paris 8) est la responsable technique (production, stockage, sauvegarde, archivage et partage des données). Concernant les entretiens et les questionnaires, la responsable légale et technique des données est Stéphanie Wojcik (CEDITEC, UPEC) (production, stockage, sauvegarde, archivage et partage des données).

L'ISC-PIF fournit le serveur (Alexandre Delanoë) de la base de données, le laboratoire Paragraphe (Paris 8) (Orélie Desfriches Doria et Waldir Lisboa Rocha) est responsable de la gestion technique.

De manière générale, le PGD sera révisé tous les 6 mois et à chaque fois que le corpus et les données vont faire l'objet d'analyse ou d'enrichissement.

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Un ingénieur de recherche a été embauché. Le temps alloué à la FAIR'isation des données est d'environ 900 heures. Il a été calculé en fonction du traitement, de la structuration et de la sécurisation des données, ainsi que de la formation de l'équipe. Mise à part la construction d'une application avec API, qui permet aux membres du projet d'avoir accès à la base de données, il s'agit également de mettre en place une stratégie d'accès public restreint, ce qui n'est pas actuellement disponible par défaut sur l'entrepôt NAKALA. Étant donné que les serveurs utilisés sont fournis par le partenaire ISC-PIF et/ou par Huma-Num, les ressources financières employées se limitent au salaire de l'ingénieur de recherche qui s'occupe de la base de données. Des ressources financières supplémentaires peuvent être envisagées si la fonctionnalité d'accès public restreint n'est pas disponible sur l'entrepôt NAKALA.

Annexe

1. Le tableau récapitulatif des commentaires collectés

Plateformes	**Twitter**	**YouTube**	**Twitch**	**Instagram**	**Total**
Nombres de commentaires	1 903 076	48 426	103 925	48 426	2 103 853