Recommandations sur l'analyse automatique de documents : acquisition, gestion, exploration

23 septembre 2019

Animatrices du groupe : Claire Nédellec, Adeline Nazarenko

Contributeurs membres du groupe : Francis André, Catherine Balivo, Béatrice Daille, Anastasia Drouot, Jorge Garcia Flores, Laurence Freyt-Caffin, Nathalie Gandon, Jacques Lafait, Nathalie Le Ba, Caroline Martin, Nathalie Morcrette, Sébastien Respingue-Perrin, Stéphanie Rennes, Pierre Zweigenbaum

Membres du groupe : Céline Delmas, Claire Lemercier, Marc Martinez, Lionel Maurel, Charline Vecco-Garda

Sommaire

Avertissement

La fouille de texte porte sur différents types de données. Les recommandations ci-après portent sur les documents protégés par un droit d'auteur, c'est-à-dire les documents écrits quels que soient leurs genres (scientifiques, journalistiques ou autres) et leur nature (tableaux, cartes, vidéos, fichiers audio, etc.). Pour les données non soumises aux droits d'auteur, il convient de se reporter au guide « Ouverture des données de recherche | Guide d'analyse du cadre juridique en France – V2 »[1].

Ces recommandations ont été rédigées par le sous-groupe de travail TDM du groupe "Développement des bonnes pratiques" du Comité pour la science ouverte. Il est composé de chercheurs en traitement automatique de la langue, en fouille de texte, de personnels d'IST, de spécialistes de plateformes et infrastructure, et d'experts juridiques. Elles sont destinées à renseigner les acteurs de la recherche publique sur les bonnes pratiques en matière de fouille de texte, suite à l'adoption de la loi française Pour une République Numérique du 7 octobre 2016 et de la directive européenne sur le droit d'auteur dans le marché unique numérique du 26 mars 2019.

[1] https://www.ouvrirlascience.fr/wp-content/uploads/2018/11/Guide_Juridique_V2.pdf

1. Acquisition des documents

Le terme **« corpus »** désigne ici un ensemble de documents constitué à des fins de fouille de texte par différents moyens. Il n'y a **pas de limitation à la taille** de ce corpus, celui-ci dépend de la nature de la recherche qui doit être faite.

L'application de logiciels de fouille de texte à un corpus de documents nécessite que les documents soient **physiquement copiés** sur la machine qui réalise les traitements. Cette « reproduction » est autorisée à certaines conditions décrites ci-dessous.

Nous ne considérons naturellement ici que les documents auxquels on accède « licitement », par exemple en bénéficiant de l'abonnement de son établissement de recherche.

1.1. Le téléchargement de documents

Le téléchargement de documents – qu'il soit manuel ou automatisé (via un robot, un *crawler* ou une interface de programmation d'application [API]) – est **réputé autorisé** s'il utilise, sans les contourner, les **moyens techniques mis à disposition** par les établissements de recherche et d'enseignement et par les fournisseurs de contenus.
Cela signifie notamment que le volume de données téléchargées sur un certain intervalle de temps doit être **raisonnable** et ne pas pénaliser les autres utilisateurs du service de téléchargement. En cas de blocage, il faut s'adresser au service IST ou à la bibliothèque qui gère l'abonnement ou directement au service de support de l'éditeur concerné.

Certains éditeurs donnent des indications[2] :

3 téléchargements par seconde sur l'API PubMed[3], ou 1 000 requêtes maximum par jour espacées de 5 secondes pour l'API PLoS Article-Level Metrics[4].

D'autres fixent des limites strictes et bloquent l'accès lorsqu'elles sont dépassées : par exemple, l'API Clarivate Web of Science[5] produit une erreur si on essaye de lancer plus d'une requête par minute.

Le téléchargement à partir des dépôts illégaux tels que Sci-Hub est naturellement interdit.

Les documents ainsi téléchargés peuvent être **conservés sans limite de temps**, et donc même après que l'abonnement qui a permis d'y accéder a été résilié.

Ce droit de téléchargement concerne exclusivement les publications licitement accessibles.

En pratique, les utilisateurs sont parfois amenés à tâtonner pour mettre au point leur procédure de téléchargement de document. Il est donc souhaitable que les éditeurs intègrent les contraintes et limites de téléchargement dans leurs API et qu'ils fournissent des explications en cas de blocage.

Ces différentes procédures de téléchargement peuvent être **combinées**. On peut ainsi :

[2] https://libguides.gc.cuny.edu/c.php?g=405353&p=4857784
[3] https://www.ncbi.nlm.nih.gov/books/NBK25497/
[4] http://api.plos.org/solr/faq/
[5] http://help.incites.clarivate.com/wosWebServicesLite/bandwidthThrottlingGroup/bandwidthThrottling.html

Modes de téléchargement des documents :

- à travers une API et avec clef d'authentification ;
- à travers un “proxy”, c’est-à-dire en configurant un navigateur pour faire des requêtes via un portail de gestion des ressources électroniques (ex. le proxy *revelec*) ;
- depuis la page d’un article sur un site auquel on a accès de façon licite pour la lecture ;
- à partir de l'identifiant ou “adresse web” (URL) du document.

- télécharger tous les articles qui ont été publiés pendant une certaine année, dans un domaine scientifique déterminé, en utilisant **différentes méthodes** pour accéder aux ressources des éditeurs des articles visés ;

- télécharger un ensemble de documents constituant un réseau de citations en exploitant un ***crawler*** auquel on fournit un ou plusieurs identifiants de documents (URL) et qui accède de proche en proche à tous les documents cités en explorant les références bibliographiques des documents déjà téléchargés ;

Exemples de reproductions autorisées :

- Lorsqu’on constitue un corpus à des fins de recherche, on peut y intégrer des sources obtenues par numérisation d’ouvrages papier.
- Pour entrainer un outil de traduction automatique sur une langue rare, on peut exploiter des textes de la BNF après les avoir numérisés par OCR (reconnaissance optique de caractères) et en extraire les données linguistiques pertinentes.

- télécharger de manière ciblée des documents depuis le site d’un éditeur auquel on a accès de façon licite en utilisant un **logiciel d’analyse de pages web et d’extraction de données** (*web scrapping*) dans le respect des limites de volume de téléchargement de l’éditeur.

1.2. Les autres modes d’acquisition

D’autres modes d’acquisition de documents numériques peuvent être envisagés à des fins de recherche, tels que la numérisation de documents papier. Ils requièrent le respect des mêmes règles concernant l’accès licite et le partage (voir paragraphe 3).

2. Utilisation de logiciels de gestion, d’exploration ou d’analyse de documents

L'utilisation d'un logiciel de **gestion, d’exploration ou d’analyse de documents** est autorisée sur les documents ou corpus obtenus licitement.

On peut ainsi utiliser un logiciel d'indexation, un logiciel de gestion bibliographique, ou tout autre logiciel – commercial ou non – de fouille de texte.

3. Partage de documents

Les documents obtenus licitement peuvent être **partagés sans limite de temps** à la condition que les bénéficiaires du partage aient les **mêmes droits d'accès que le premier accédant**, et ceci pour chacun des documents partagés.

Exemples de bénéficiaires du partage :

- les collaborateurs d'une même équipe de recherche ;
- les collaborateurs d'un projet appartenant à des organismes de recherche ayant souscrit les mêmes abonnements.

Les bénéficiaires peuvent partager les corpus téléchargés ou les documents une fois prétraités pour les opérations de TDM, qui sont en général des produits dérivés (voir paragraphe 3.2).

3.1. Précautions pratiques

Dans le cas de corpus composés de documents sous licences différentes, il faut **vérifier, pour chaque source**, que les bénéficiaires du partage ont un **accès licite**. Il convient également d'informer les bénéficiaires du partage de ces règles de diffusion. Le partage hors ayant droit s'apparente à du **piratage**. Les correspondants IST, bibliothécaires et documentalistes des établissements peuvent être consultés pour vérifier les droits d'accès.

Dans le cas de documents obtenus **hors abonnement** faisant l'objet d'une **convention particulière** associée par exemple à un projet, **c'est la convention qui fixe les conditions de partage**. Selon le degré de confidentialité, les contraintes décrites par la convention peuvent être très fortes et l'accès limité à des personnes nommément désignées.

Exemples de documents régis par des conventions spécifiques :

- ensemble documentaire appartenant à un partenaire de projet (rapports techniques ou stratégiques, rapports d'incident, etc.) ;
- entretiens non anonymisés ;
- rapports cliniques.

Afin d'éviter les copies illicites, la **conservation des documents** doit répondre aux contraintes de sécurité en vigueur pour le type de sources considérées. On peut par exemple restreindre les droits d'accès par authentification, utiliser pour le stockage un ordinateur isolé du réseau, etc.

3.2. Produits dérivés

Les produits dérivés **résultent en général d'un prétraitement** destiné à faciliter des étapes ultérieures d'analyse et d'exploration documentaire. Ils sont **assimilés à des copies des documents source** : le format, la structure ou le contenu peuvent être modifiés, mais il est possible de reconstituer le texte source ou une partie substantielle du texte source.

Exemples de documents ou produits dérivés :

- copies d'articles scientifiques au format texte après suppression des marques de styles ;
- traduction de documents ;
- textes enrichis d'annotations (étiquetés en parties du discours, par exemple) ;
- extraits de documents correspondant à des parties significatives du texte d'origine, comme l'introduction et la conclusion par exemple ;
- documents lemmatisés où chaque occurrence de mot (ex. "finissez") est remplacé par son lemme ou forme canonique ("finir") ;
- documents simplifiés par la suppression des mots grammaticaux ;
- index d'un document associant à chaque mot du vocabulaire la position de ses occurrences dans le texte source.

Les produits dérivés de documents sont **partageables aux mêmes conditions que les documents dont ils dérivent**. En cas de doute et pour distinguer un produit dérivé des données de la recherche résultant de l'exploration ou de l'analyse de documents (voir paragraphe 4), il convient de consulter un expert en TDM qui pourra déterminer s'il est possible ou non de reconstituer automatiquement une partie substantielle du texte source.

3.3. Cas particuliers

Les conditions de partage des **documents dits "publics" ou "ouverts"** sont à analyser au cas par cas. Il faut se reporter aux **licences des documents** avant de les diffuser ou de les rendre publics.

Pour la **licence de type Creative Commons BY** – la plus courante –, il faut créditer l'Œuvre, c'est-à-dire intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Œuvre quand cela est autorisé. **La licence CC-BY-ND** (*No derivative*) **interdit le partage de la version modifiée**.

Les résumés et les métadonnées bibliographiques des articles qui sont fréquemment disponibles sur des sites de base bibliographique ne sont en général pas diffusables sans condition ; ni publiquement, ni entre chercheurs ne bénéficiant pas des mêmes accès. **Les conditions doivent être vérifiées avant partage**.

Dans le cas particulier de **l'organisation de compétitions** (*shared task*), le partage du corpus de la compétition est possible, au-delà des accès licites, pour un **nombre limité de documents** ou de parties de documents **justifiés par l'objet** de la compétition. L'accès est en général restreint à des utilisateurs identifiés. L'usage est limité à l'entraînement et au test des outils de fouille de texte dans le cadre de la compétition. Le partage **peut être sans limite de temps** pour favoriser la reproductibilité et la comparaison. Quand cela est possible, il est préférable d'organiser des compétitions portant sur des documents avec des licences ouvertes. Dans tous les cas, les auteurs doivent être cités.

Exemples de corpus de compétition :

- les références de la base PubMed sont rediffusables en ligne à la condition de citer la source ;
- la compétition National NLP Clinical Challenges (n2c2) utilise un corpus de rapports médicaux. L'accès et l'utilisation du corpus est soumis à la signature d'un accord de confidentialité très strict et nominatif.

4. Exploitation des résultats de l'exploration ou de l'analyse de documents

Les résultats de l'exploration ou de l'analyse de documents sont qualifiés de **données de la recherche s'ils ne contiennent pas de partie significative des documents**. C'est ce qui les distingue des produits dérivés (voir paragraphe 3.2).

En pratique, la majorité des algorithmes d'analyse et d'exploration de documents transforme et « déconstruit » les textes sources de manière irréversible.

Exemples de résultats constituant des données de la recherche :

- une concordance établie à partir d'un ouvrage (liste des contextes d'occurrence d'un mot ou d'une expression clé) dès lors que la taille du contexte pris en compte reste modeste au regard de la taille de l'œuvre :
- une classification de documents associant les DOI à des catégories, mais ne donnant pas accès au contenu des documents eux-mêmes ;
- un outil de traduction entraîné à partir d'un corpus bilingue ;
- une grammaire ou un analyseur syntaxique construit à partir d'un large corpus ;
- un plongement de mots (ou « word embeddings », dictionnaire associant à chaque mot une représentation dans un espace vectoriel) obtenu par analyse d'un corpus représentatif d'une langue ou d'un domaine spécifique ;
- le résumé d'un document obtenu par des moyens automatiques dès lors que le résumé ne contient qu'un nombre modeste d'extraits de taille modeste, sachant que le nombre d'extraits et la taille de ces extraits sont à apprécier au regard de la taille du document source (le résumé ne doit pas comporter de partie significative de l'œuvre originale, voir paragraphe 5).

Ces données de la recherche sont **soumises aux mêmes règles que les autres données de la recherche** et il convient de se reporter aux règles des organismes de recherche ou aux conventions particulières s'appliquant sur les données considérées, dont la RGPD (voir *Le guide sur l'ouverture des données de recherche*[6]). **L'utilisation de ces données peut être commerciale ou non commerciale.**

Un cas particulier assez fréquent concerne les **données obtenues par annotation de textes sources**. Au-delà des métadonnées qui sont généralement associées globalement à un document pris dans son ensemble, on peut ajouter des étiquettes aux mots, groupes de mots, phrases, paragraphes, etc., qui composent le document.

Exemples de listes d'annotations :

- une séquence de parties du discours ou d'étiquettes morphosyntaxiques obtenue par annotation d'un corpus de texte du XVIe siècle ;
- la séquence de mots-clés ou d'entités nommées extraits d'articles de presse ;
- la liste des thèmes abordés dans un rapport scientifique ;
- la liste des lieux et personnes mentionnées.

Sauf dans les cas rares et très particuliers où l'annotation permettrait de reconstituer le texte source, **cette liste d'annotations constitue une donnée de la recherche** et peut être partagée ou exploitée indépendamment du texte source à partir duquel elle a été construite.

5. Diffusion d'extraits de documents

Les extraits des documents contenus dans les résultats de l'exploration ou de l'analyse de documents sont diffusables publiquement soit **si la licence le permet**, soit si ce sont des **parties non significatives de ces documents** (de courtes citations).

Dans tous les cas, **la référence du document doit être précisée** (auteur, éditeur, titre de l'ouvrage, date d'édition, etc.). La diffusion de larges extraits ou des

Exemples d'extraits non diffusables :

- les figures ou dessins sont à considérer comme des extraits significatifs quand ils sont des œuvres originales ;
- il est interdit – à moins que la licence ne l'autorise explicitement – de diffuser publiquement pour l'organisation de compétitions des articles entiers ou des portions importantes comme les introductions ou conclusions des articles.

[6] https://www.ouvrirlascience.fr/wp-content/uploads/2018/11/Guide_Juridique_V2.pdf

œuvres complètes est interdite par défaut, à moins que la licence ne l'autorise explicitement (exemple : *Creative Commons*).

Il n'y a pas de règles précises pour apprécier ce qu'est une partie non significative de document, mais **la taille des extraits doit être justifiée** par les nécessités de vérification et de contextualisation des résultats de l'analyse et de l'exploration de documents.

La diffusion d'extraits ne doit pas contourner la restriction sur la diffusion des œuvres : **il ne faudrait pas que les extraits diffusés dispensent les lecteurs de recourir à l'œuvre originale**.

Exemples d'extraits diffusables :

- les extraits de trois ou quatre lignes publiés par Google, ou Google Scholar sont autorisés par l'usage ;
- sur un article scientifique d'une vingtaine de pages, on peut faire une citation d'une dizaine de lignes ;
- la liste des "entités nommées", (les noms de ville, noms de personnes, etc.) citées avec leur contexte peut être diffusée pour permettre la vérification de la pertinence de l'extraction des entités ;
- pour la visualisation des relations ou l'analyse des coréférences entre entités, il peut être justifié de diffuser un paragraphe pour préserver la cohérence du propos ;
- dans le cas du résumé automatique par extraction et concaténation de phrases, la diffusion est autorisée car on peut considérer que chaque phrase est un extrait court mais il faut que le nombre de phrases extraites d'un même document soit modeste au regard de la taille du document.