

Baromètre français de la Science Ouverte

Données, code et logiciels : une méthodologie innovante

L'analyse du texte intégral des publications est une méthode générique pour détecter l'utilisation, la production et le partage de jeux de données et de logiciels.

Première étape

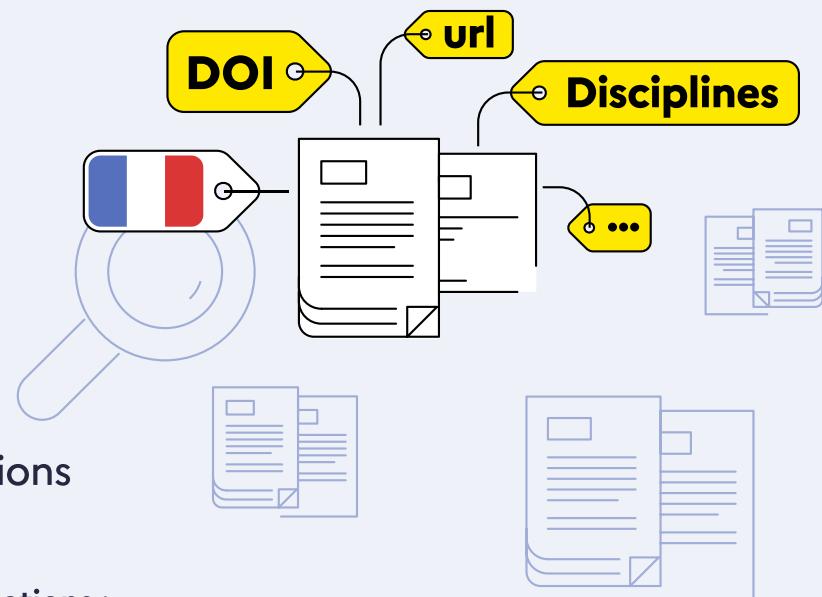
Point de départ : répertorier et caractériser la production scientifique française

Récupération des métadonnées ouvertes enrichies par le Baromètre portant sur les publications (DOI, URL, disciplines, affiliations ...)

Découvrir la méthodologie du baromètre portant sur les publications :

barometredelascienceouverte.esr.gouv.fr/a-propos/methodologie#publications

et le preprint associé hal.science/hal-03651518



Deuxième étape

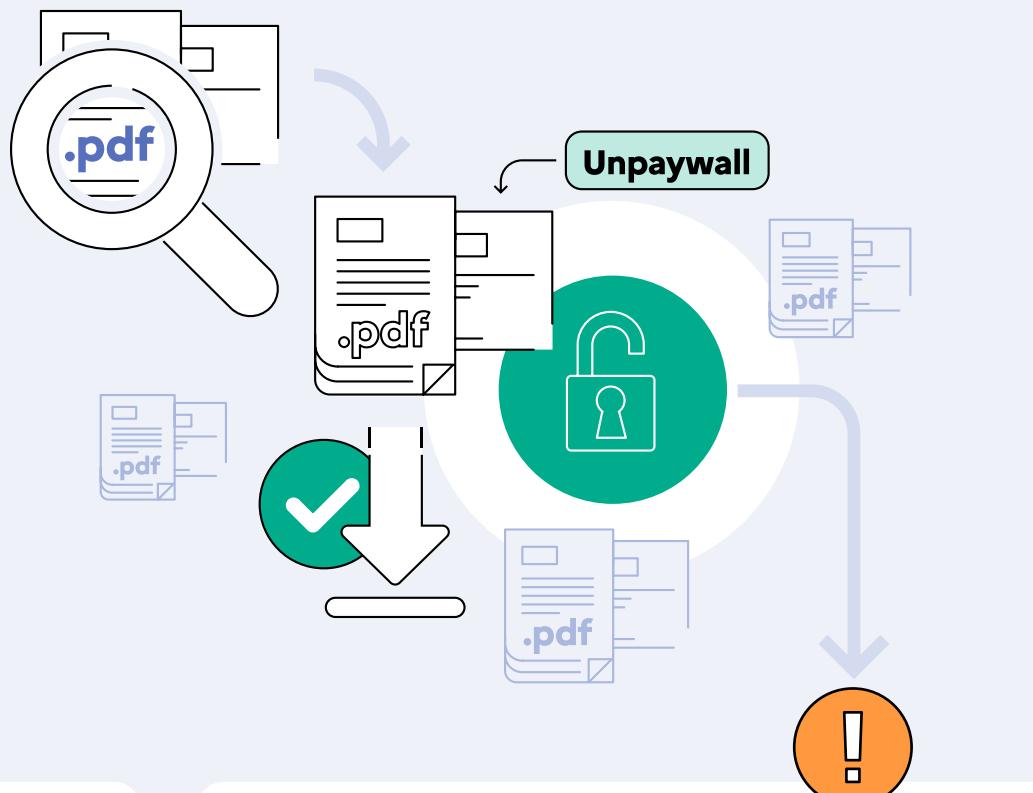
Télécharger le texte intégral de toutes les publications du corpus

Publications en accès ouvert

Pour chaque publication, grâce à son DOI, Unpaywall fournit des URL qui permettent de télécharger une version du texte intégral au format PDF.



des publications en accès ouvert du baromètre ont ainsi été téléchargées



Quelques difficultés techniques à souligner : redirection, page inexistante, site momentanément indisponible, lien brisé, blocage des robots, publication en accès libre dans le navigateur mais non téléchargeable automatiquement...



Publications en accès fermé

Utilisation des abonnements des partenaires du projet (abonnement national, abonnements de l'Université de Lorraine).



L'ordonnance de transposition de la directive européenne 2019/790* sur le droit d'auteur et les droits voisins dans le marché unique numérique le prévoit.

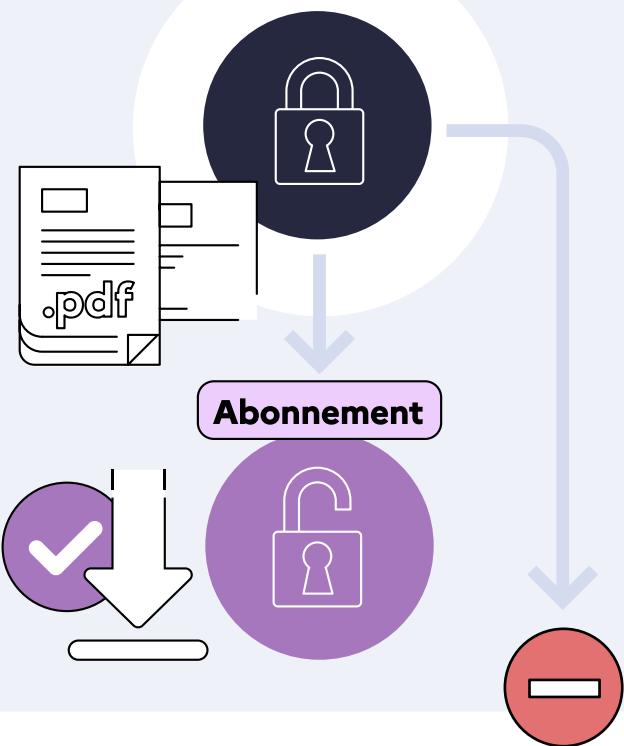
* <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A32019L0790>

38%

des publications en accès fermé du baromètre ont ainsi été téléchargées

Mais certains éditeurs avec lesquels nous avions un contrat d'abonnement ont mis en œuvre **des dispositifs contraignants pour encadrer ces téléchargements (API, jetons de téléchargement...)**.

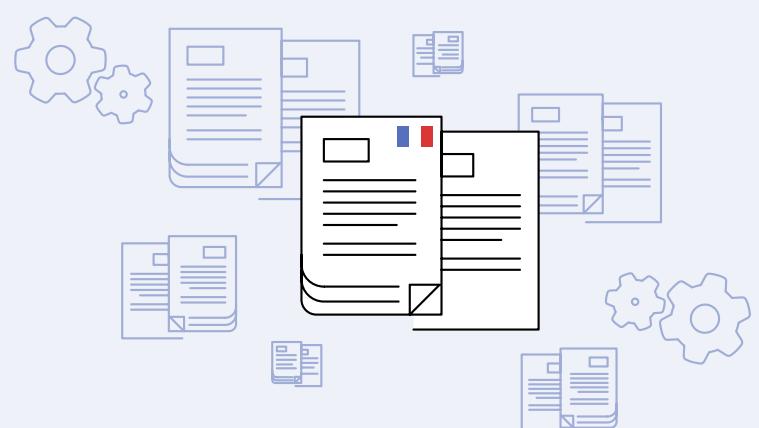
Pour les éditeurs pour lesquels nous ne disposions pas d'abonnement, il n'a pas été possible de télécharger les PDF en accès fermé.



Troisième étape

Un processus d'enrichissement reposant sur de l'apprentissage automatique

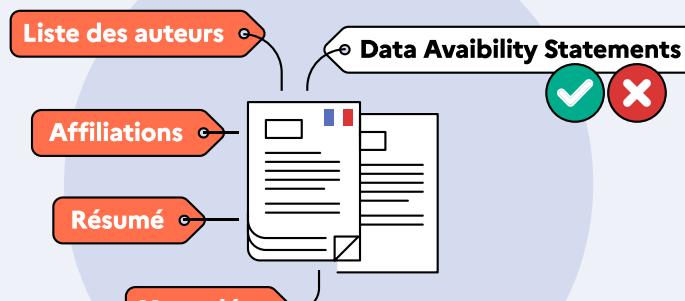
Pour ces PDF téléchargés, **3 outils - fonctionnant notamment sur le Deep Learning - ont été utilisés :**



GROBID

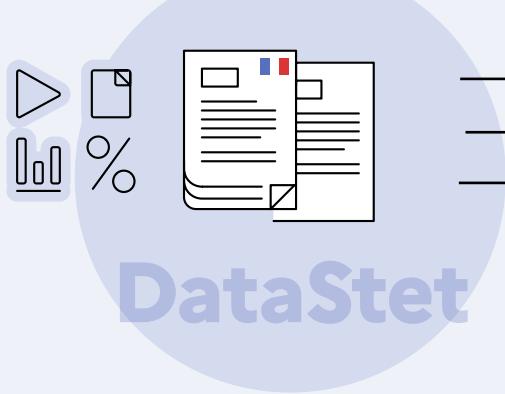
L'outil extrait les métadonnées et structure le texte intégral de la publication dans un format standard XML-TEI :

Dans le cadre de ce projet, la détection de "Data Availability Statements", ou **déclarations sur la mise à disposition des données**, a été ajoutée à GROBID.



DataStet

Il détecte, dans le texte intégral, toutes les mentions de jeux de données



Chaque mention est caractérisée selon 3 catégories qui peuvent se cumuler :

- Utilisation
- Production
- Partage

Softcrite

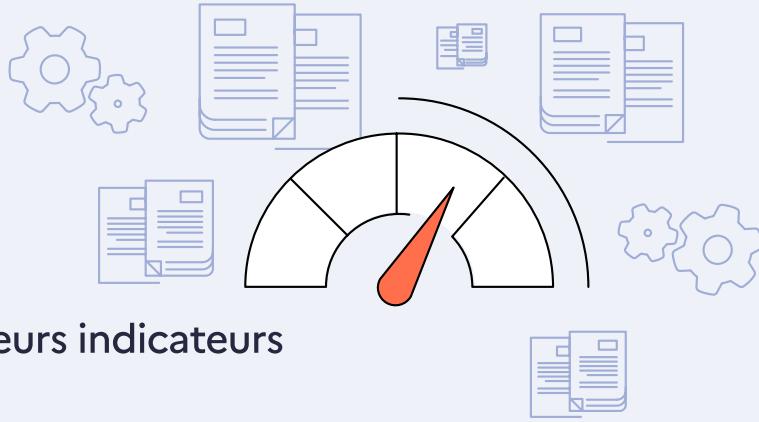
Il détecte, dans le texte intégral, toutes les mentions de code et logiciels



Quatrième étape

La production d'indicateurs issus de ces calculs et enrichissements

Pour les jeux de données et les codes et logiciels, plusieurs indicateurs ont été produits selon une **analyse en entonnoir** :



Pour les données de la recherche avec **DataStet** :

Parmi les **publications analysées**,

part de celles qui mentionnent l'utilisation de données dans le texte intégral

Parmi celles qui mentionnent l'utilisation de données,

part de celles qui mentionnent la production de leurs données

Parmi celles qui mentionnent la production de leurs données,

part de celles qui mentionnent le partage de leurs données

Softcite

Une analyse identique est menée concernant les code et logiciels.

GROBID

Part de publications qui incluent une section "Data Availability Statement"

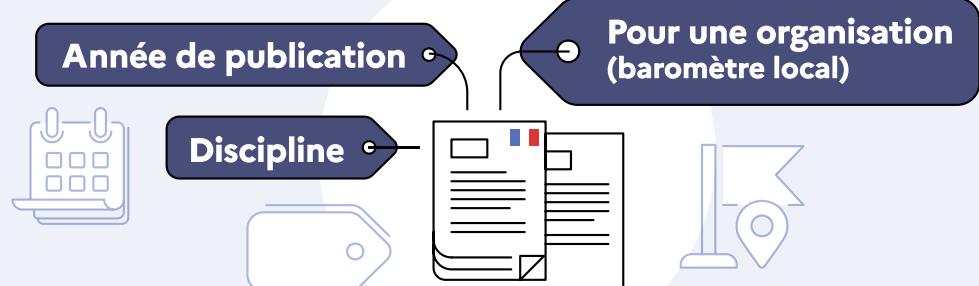
(déclaration sur la mise à disposition des données).



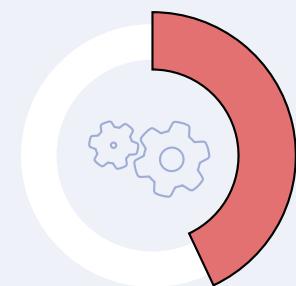
Ces indicateurs globaux sont déclinés selon des dimensions propres à la publication :



Les publications inaccessibles et les coûts des traitements ont conduit à limiter l'analyse à **43% des publications**.



Elle a été pour le moment partiellement menée.



Méthodologie publiée sur HAL

hal.science/hal-04121339

Cette nouvelle brique méthodologique a été réalisée avec :

Grâce au soutien de :