
DMP du projet "Baromètre Science Ouverte Données et codes"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR).

Renseignements sur le plan

Titre du plan	DMP du projet "Baromètre Science Ouverte Données et codes"
Domaines de recherche (selon classification de l'OCDE)	
Langue	fra
Date de création	2022-05-02
Date de dernière modification	2022-09-14
Documents (publications, rapports, brevets, plan expérimental...), sites web associés	<ul style="list-style-type: none">• Baromètre français de la Science Ouverte : https://barometredelascienceouverte.esr.gouv.fr/

Renseignements sur le projet

Titre du projet	Baromètre Science Ouverte Données et codes
Acronyme	BSO3 ou BSO Données et codes
Résumé	Création de nouveaux indicateurs, portant sur les données de la recherche et les codes logiciels, afin de mesurer la mise en œuvre d'une politique publique de la donnée.
Sources de financement	<ul style="list-style-type: none">• France Relance :
Date de début	2021-09-22
Partenaires	<ul style="list-style-type: none">• University of Lorraine ()• Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation ()• Inria ()

Produits de recherche :

1. Default research output

Contributeurs

Nom	Affiliation	Rôles
Bassinot Aricia	Université de Lorraine	
Bracco Laetitia		<ul style="list-style-type: none"> • Coordinateur du projet • Personne contact pour les données • Responsable du plan de gestion de données
Jeangirard Eric	Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation	
L'Hôte Anne	Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation	
Lopez Patrice	science-miner	
Lutz Jean-François	Université de Lorraine	
Romary Laurent	Inria	
Weisenburger Emmanuel	Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation	

Droits d'auteur :

Le(s) créateur(s) de ce plan accepte(nt) que tout ou partie de texte de ce plan soit réutilisé et personnalisé si nécessaire pour un autre plan. Vous n'avez pas besoin de citer le(s) créateur(s) en tant que source. L'utilisation de toute partie de texte de ce plan n'implique pas que le(s) créateur(s) soutien(nen)t ou aient une quelconque relation avec votre projet ou votre soumission.

DMP du projet "Baromètre Science Ouverte Données et codes"

1. Description des données et collecte ou réutilisation de données existantes

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

Le Baromètre Science Ouverte Données et codes réutilise les données du [Baromètre français de la Science Ouverte](#) seconde génération, notamment le corpus de métadonnées de publications. Ces données sont mises à disposition sur [data ESR](#) sous licence ouverte Etalab.

De nouvelles données seront générées par deux biais :

- L'analyse de publications en texte intégral pour y détecter des mentions de jeux de données et de logiciels,
- Le moissonnage des métadonnées d'entrepôts de jeux de données pour déterminer le corpus des données de recherche françaises.

L'analyse des publications s'appuie sur des modèles d'apprentissage automatique eux même entraînés sur des données annotées manuellement provenant de corpus existant et d'annotations propres au projet.

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

Les données sont de deux types : résultats de l'analyse de corpus à grande échelle des publications françaises visant à produire les indicateurs sur l'accès libre et corpus annotés manuellement visant à entraîner les modèles d'apprentissage automatique impliqués dans l'analyse.

Les données seront produites selon plusieurs formats :

- des fichiers en jsonl et en csv pour les résultats d'extraction des jeux de données et de codes logiciel ;
 - des fichier TEI XML et JSON pour les corpus annotés manuellement dédiés à l'apprentissage automatique.
-

2. Documentation et qualité des données

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

Relativement aux résultats d'extraction à grande échelle sur les publications :

- aucun standard de métadonnée particulier ne sera utilisé. En revanche, les bonnes pratiques en matière de code seront respectées pour les données également : présence d'un fichier README, versionnage des fichiers, publication d'une méthodologie.
- à terme, une documentation d'accompagnement sera produite pour accompagner les établissements dans leur appropriation du Baromètre et favoriser les déclinaisons locales.

Relativement aux corpus annotés destinés l'apprentissage automatique :

- les données réutiliseront les formats et métadonnées existants basés sur la TEI XML,
 - les données sont ouvertes et partagées sur Zenodo, ce qui implique la conformité des métadonnées du jeu de données et de la gestion de version de cette plateforme.
-

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

Pour la détection de mentions de jeux de données et de logiciels, la mesure de F1-score cible est fixé à 80%. Cette mesure de qualité suit les standards actuels d'évaluation en terme d'extraction d'entités. En particulier, chaque modèle est évalué :

- d'une part par "10-folds cross-validation" (moyenne de 10 partitions entraînement/évaluation du corpus annoté),
- d'autre part par "holdout set" stable, approche habituellement favorisée par les spécialistes en apprentissage automatique pour sa

fiabilité par rapport au biais d'apprentissage.

Relativement à l'annotation manuelle, nous visons une qualité d'annotations de type "gold", reposant sur une annotation en double aveugle de chaque document, suivi d'une étape de réconciliation sur les cas de désaccord entre annotateurs. Il s'agit du standard de qualité le plus élevé en terme d'annotation manuelle de corpus.

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 ?

Le code est stocké sur Github. Les données produites font l'objet d'un stockage sur le cloud privé OVH Object Storage, avec une mise à jour régulière.

Les corpus annotés étant des extensions de corpus existants, ils feront l'objet de nouvelles versions de corpus déposés sur Zenodo et référencés via DataCite.

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

Les données sensibles, à savoir les publications en texte intégral, ne seront pas disponibles publiquement. Elles ne seront accessibles qu'aux membres institutionnels du projet, via le cloud privé OVH Object Storage.

Les données des corpus annotés manuellement sont entièrement dérivées de documents disponibles en licence CC-BY, et ne sont donc pas sensibles en terme de copyrights.

4. Exigences légales et éthiques, 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 ?

Aucune donnée à caractère personnel n'est traitée.

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 ?

Les titulaires du copyright sur les données réalisées dans le projet sont les institutions impliquées dans le projet et produisant ces données.

Les corpus de métadonnées sont destinés à être ouverts en licence Etalab.

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

Il n'y a pas de question éthique dans ce projet.

5. Partage des données et conservation à 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 ?

Le code est partagé via le dépôt <https://github.com/Barometre-de-la-Science-Ouverte/>

Les corpus seront déposés au fur et à mesure de l'avancement du projet. Il n'y a pas de restriction au partage du code et des corpus de métadonnées au delà des licences indiquées précédemment. En revanche, les publications en texte intégral (PDF) ne seront pas redistribuées et n'auront qu'un usage interne au projet.

A la fin du projet, les données ainsi produites seront hébergées par le Ministère et indexées sur [son propre portail d'Open Data](#) sous licence ouverte v2.0 (Etalab), sur le même modèle que les données du [Baromètre sur les publications](#).

Un dépôt au minimum annuel des données du projet sera également effectué dans [Recherche Data Gouv](#). Ce dépôt comprendra les données du Baromètre publications (BSO2) et du Baromètre données et codes (BSO3).

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) ?

A la fin du projet, le code et les corpus de métadonnées dans leur ensemble feront l'objet d'un dépôt dans HAL et dans Software Heritage.

Les corpus annotés manuellement seront déposés sur Zenodo, sous la forme de nouvelles versions de corpus existants ayant été étendus.

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

Le code et les corpus de métadonnées sont produits dans des formats libres. Aucun logiciel propriétaire n'est nécessaire pour y accéder et pour les utiliser.

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

Un identifiant Software Heritage sera attribué au code à l'occasion du dépôt conjoint HAL / Software Heritage.

Les jeux de données déposés dans Recherche Data Gouv se verront attribuer un DOI.

Des DOI relatifs aux versions des corpus annotés manuellement seront attribués lors des dépôts sur Zenodo.

6. Responsabilités et ressources en matière de gestion des données

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) ?

Le gestionnaire du code et des corpus de métadonnées est l'équipe du Ministère de l'Enseignement Supérieur et de la Recherche (MESR).

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) ?

Il n'y a pas de ressource spécifique dédiée hormis le temps de travail des membres de l'équipe. Par le biais de l'utilisation de formats et d'outils libres, les données seront FAIR by design.