Plan de gestion de données du projet "Baromètre Science Ouverte Données et codes"

Aricia Bassinet, Laetitia Bracco, Eric Jeangirard, Anne L'Hôte, Patrice Lopez, Jean-François Lutz, Laurent Romary, Emmanuel Weisenburger

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	• 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	• France Relance :
Date de début	2021-09-22
Partenaires	• University of Lorraine 04vfs2w97 • Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation • Inria 02kvxyf05

Produits de recherche :

1. Default research output (Jeu de données)

Contributeurs

Nom	Affiliation	Rôles
Bassinet Aricia	Université de Lorraine	
Bracco Laetitia		• 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	

DMP du projet "Baromètre Science Ouverte Données et codes"

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

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.

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

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.

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

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.

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

Aucune donnée à caractère personnel n'est traitée.

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.

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

---

5. Partage des données et conservation à long terme

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).

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.

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.

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

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

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.