Données ouvertes au centre des IAG

Biennale 2023, Enssib

Ghislaine CHARTRON

Professeur du CNAM

le cnam

Données au centre des IAG…

- Pour les algorithmes d'apprentissage
- Couverture de BARD et ChatGPT ?...
- Données annotées (structurées) essentielles pour l'apprentissage supervisé
- Des pratiques illicites ou non consenties (*Book3* pour les livres, scrapping des media…)

Protéger ses données contre les IAG

- Utiliser *l'Opt-out* introduite dans la directive droit d'auteurs 2019 concernant les usages « non recherche » (ex: Clause type du SNE)

- Bénéfice: valorisation des données par des IAG internes *(ex: Getty Image, Lexis-Nexis, X-Tweeter avec l'IAG Grok)*

- Risques :
  * les IAG grand-public nourries avec des données de faible qualité.
  → enjeux politiques pour les sociétés, désinformation, manipulation.
  * devenir peu visible, impact réduit des ses données

Ouvrir ses données aux IAG

- Bénéfice :
  * Contribuer à la qualité des services d'information automatisés grand-public, réduire les stratégies de désinformation
  * Plus de citations

- Risques :
  *Détournement de la valeur des producteurs de contenus
  *La science ouverte au profit des acteurs technologiques les plus puissants (Google, Open-AI…)

Dilemme…Vers une nouvelle régulation ?

- Entre les IAG et les media ?
- Partage de la valeur, analogie avec la création en 2019 d'un droit voisin pour la presse face à Google news ?

N° 1630

ASSEMBLÉE NATIONALE

CONSTITUTION DU 4 OCTOBRE 1958

SEIZIÈME LÉGISLATURE

Enregistré à la Présidence de l'Assemblée nationale le 12 septembre 2023.

PROPOSITION DE LOI

visant à **encadrer** *l'***intelligence artificielle** *par le* **droit** *d'***auteur**,

(Renvoyée à la commission des affaires culturelles et de l'éducation, à défaut de constitution d'une commission spéciale dans les délais prévus par les articles 30 et 31 du Règlement.),

présentée par Mesdames et Messieurs

Guillaume VUILLETET, Claire PITOLLAT, Olga GIVERNET, Dominique

Améliorer la transparence et le partage des données d'apprentissage

https://www.dataprovenance.org/
https://github.com/Data-Provenance-Initiative

Explorateur de provenance des données

La Data Provenance Initiative est un audit à grande échelle d'ensembles de données d'IA utilisés pour former de grands modèles de langage. Dans un premier temps, nous avons retracé plus de 1 800 ensembles de données de réglage texte à texte populaires, depuis leur origine jusqu'à leur création, en cataloguant leurs sources de données, licences, créateurs et autres métadonnées, pour que les chercheurs puissent les explorer à l'aide de cet outil. Le but de ce travail est d'améliorer la transparence, la documentation et l'utilisation éclairée des ensembles de données en IA.

Vous pouvez télécharger ces données (avec des filtres) directement depuis la collection de provenance des données .

Si vous souhaitez contribuer ou discuter, n'hésitez pas à contacter les organisateurs à data.provenance.init@gmail.com .

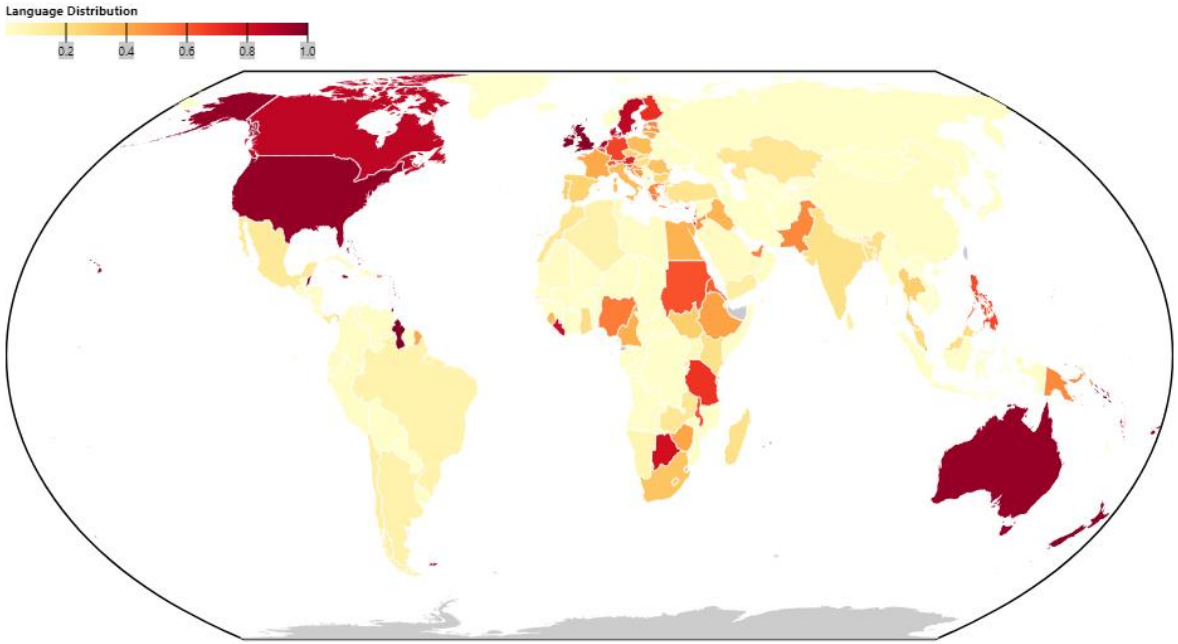
NB : Il est important de noter que nous collectons *les licences autodéclarées* , auprès des journaux et référentiels qui ont publié ces ensembles de données, et que nous les catégorisons selon nos meilleurs efforts, dans le cadre d'une initiative de recherche volontaire et de transparence. Les informations fournies par nos travaux et les résultats de la Data Provenance Initiative ne constituent **PAS** et ne sont **PAS** destinés à constituer des conseils juridiques ; au lieu de cela, toutes les informations, contenus et documents sont uniquement destinés à des fins d'information générale.

Dépôt de données Papier

Representation linguistique par pays

Nous visualisons d'abord la couverture linguistique par pays, en fonction des langues parlées et de leur représentation dans la collection de provenance des données. dans le pays$k$qui parlent une langueje, et$w_{je\ je}$qui est un indicateur binaire qui vaut 1 si l'ensemble de données$je \in D$contient une languejeet 0 sinon.

$$S_k = \sum_{l \in L} \left(p_{.kl\_} \times \sum_{je \in r\acute{e}} w_{je\ je} \right)$$



Qqs références

LINC-CNIL, [Dossier IA générative] - ChatGPT : un beau parleur bien entraîné, avril 2023, https://linc.cnil.fr/dossier-ia-generative-chatgpt-un-beau-parleur-bien-entraine

Longpre, S., “The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI”, arXiv e-prints, 2023. doi:10.48550/arXiv.2310.16787 .

Nikos Smyrnaios et Charis Papaevangelou, « Réguler la dépendance », *Balisages* [En ligne], 6 | 2023, mis en ligne le 21 septembre 2023, consulté le 07 novembre 2023. URL : https://publications-prairial.fr/balisages/index.php?id=1055

Actus IA, https://www.actuia.com/

Chartron, G. & Broudoux, É. (2015). Enjeux géopolitiques des données, asymétries déterminantes. Dans : Évelyne Broudoux éd., *Big Data - Open Data : Quelles valeurs ? Quels enjeux: Actes du colloque « Document numérique et société », Rabat, 2015* (pp. 65-83). Louvain-la-Neuve: De Boeck Supérieur. https://doi.org/10.3917/dbu.chron.2015.01.0065