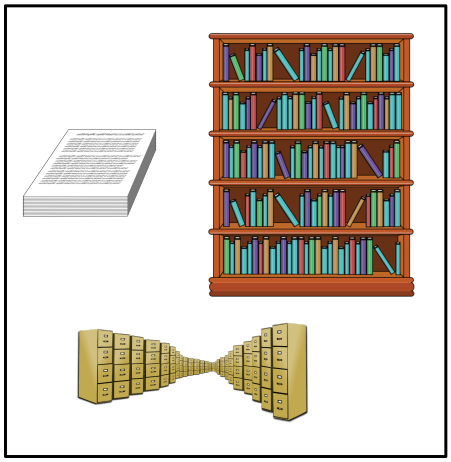
La plateforme OLKi

Pierre-Antoine Rault, Aurore Coince, Christophe Cerisara

Types de données de recherche



Note: 2 grandes classes de données:

- les données anciennes, figées: ex les benchmarks...
  - CINES, ORTOLANG sont parfaits
- les données en construction: projet ANR, thèse, conférence, un hackaton...
  - échangées par email, déposées sur un site web, ResearchGate, academic torrents...

Emily M. Bender @emilymbender · 4 Nov 2017

I don't think he does -- at least not judging from what's in Wang & Eisner 2016 (p.491):

> As a result, it is challenging to develop systems that will discover structure in new languages in the same way that an image segmentation method, for example, will discover structure in new images. The limited resources even make it challenging to develop methods that handle new languages by unsupervised, semi-supervised, or transfer learning. Some such projects evaluate their methods on new sentences of the same languages that were used to develop the methods in the first place—which leaves one worried that the methods may be inadvertently tuned to the development languages and may not be able to discover correct structure in other languages. Other projects take care to hold out languages for evaluation (Spitkovsky, 2013; Cotterell et al., 2015), but then are left with only a few development languages on which to experiment with different unsupervised methods and their hyperparameters.

2

Hal Daumé III @haldaume3 · 4 Nov 2017

the whole "the distribution of langs I care about is distribution of those that people choose to make treebanks of" doesn't compute for me >

2 1

Hal Daumé III @haldaume3 · 4 Nov 2017

furthermore arg doesn't work even if you believ that: samples are DEFINITELY not IID (less likely to get TB of lang that already has a TB)

1 2

(((J()(J() 'yoav)))) @yoavgo · 4 Nov 2017

I think this is mostly about NLP people meaning something completely different than Linguists when they say "language independent". >

Note: Type data: articles:

- en TAL/SDL, fréquent de travailler sur des bases d'articles scientifiques;
- ex: méta-analyse de papiers sur Twitter; données annotées appartiennent à Twitter !
- "fuites" de données importantes; pourquoi ? Manque de fonctionalités réseau social scientifique libre

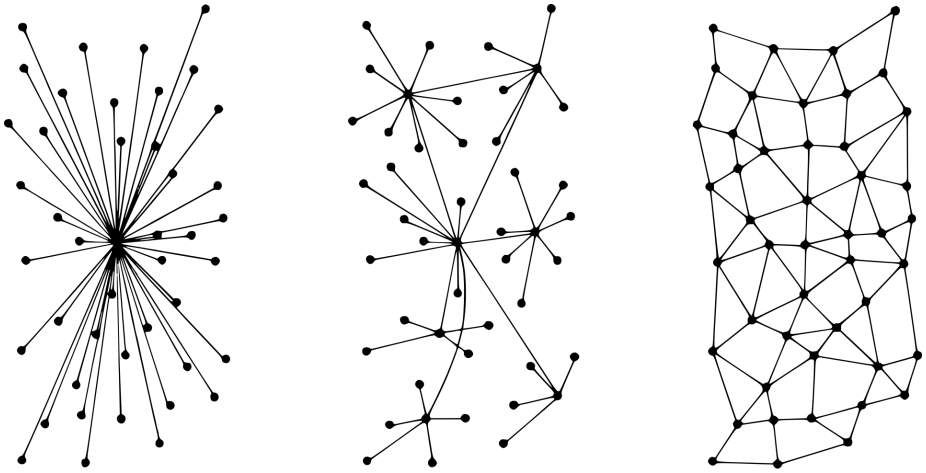
---

Le projet OLKi

Note: OLKi vise à combler ce manque:

- solutions pour distributer / échanger les données en cours de création
- interactions via les réseaux sociaux
- complémentaire des centres de resources comme ORTOLANG

Silo vs. fédéré vs. P2P

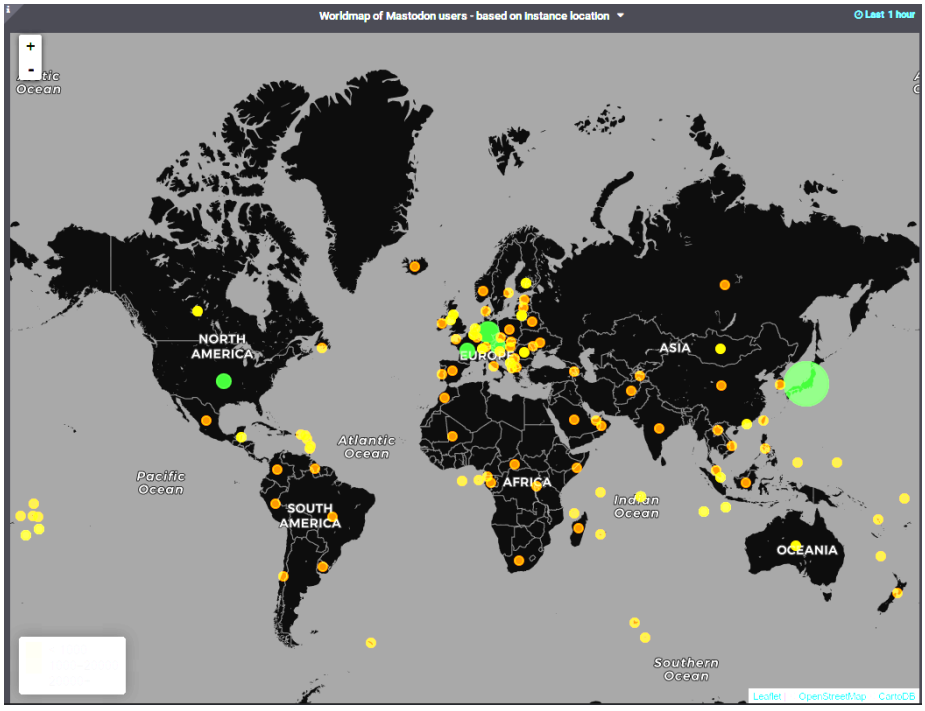


ActivityPub

Note: La manière de distribuer les données est importante

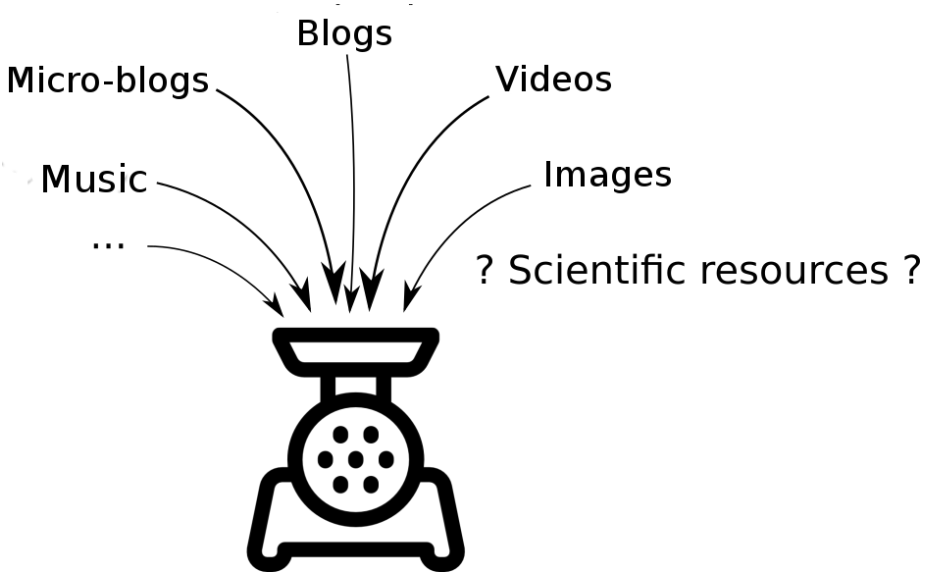
- gauche: silos centralisés: facile d'accès, mais contrôle centralisé, coût élevé
- droite: pair à pair, peu efficace
- approche fédérée, qui favorise les communautés: adoptée par OLKi, implémentée via un protocole récent du W3C

Le Fediverse



Note: Protocole déjà utilisé par 2M de personnes dans le Fediverse Très présent Europe + Japon: réseau social libre, éthique, ouvert, maître de ses données. OLKi fait partie du Fediverse, et bénéficie des outils et de la communauté

La modalité “corpus”



Note: - OLKi enrichit le Fediverse en ajoutant une nouvelle modalité d’échange pour les données - Fait le lien entre ces réseaux citoyens et les réseaux de recherche académiques

---

Fonctionnalités

- Déployer une instance: docker
- Uploader un corpus local
- Downloader corpus distant (OAI-PMH)
- Fil de discussion fédéré (ActivityPub)
- Suivre une autre instance (ActivityPub)

Note: Fonctionnalités aujourd’hui

---

https://olki-social.loria.fr

Note: plateforme de démonstration héberge des corpus gère users et communauté fédération des corpus fédération des commentaires avec Mastodon

---

En résumé

- éthique: producteurs de données au contrôle
- ouverte: code source + API
- coûts partagés, passe à l'échelle
- résistante aux pannes
- Complémentaire des silos institutionnels
- Ouverte à / pour l'expérimentation

Data: partager et traiter

- **Federated Learning**: la vision Google / IBM / WeBank

Note: base de travail => aller plus loin Exploiter les corpus + bénéficier du réseau FedDL = problèmes majeurs: secret / privauté

Decentralized Federated Learning

Vision

- Apprentissage machine **local**
- Partager les **modèles** sur OLKi
- Les fusionner (multi-task transfer learning)

Note: partage automatique de communauté en communauté

Decentralized Federated Learning

Intérêt

- Protéger les données privées ou secrètes (industrielles)
- Des IA à taille humaine, collaboratives

Note: protège mieux que en centralisé: modèles plus durs à localiser, partiels

---

Cours en ligne

- Héberge des transparents de cours:
  - synchronisés sur un Master
  - annotations live synchronisées
  - commentaires audio pour du offline
- Intéractions via réseaux sociaux libres

---

olki.loria.fr pour...

- Partager des données, garder le contrôle
- Interagir sur les réseaux sociaux libres
- Bâtir des communautés
- Lier scientifiques et citoyens
- Initier des IA partagées