

Pour des IA francophones

Open Lab - 3 octobre 2019

piaf@data.gouv.fr



Piaf

etalab gouv.fr

Programme

9h30 - Accueil des participants

9h45 - Présentation du projet PIAF et de ses acteurs

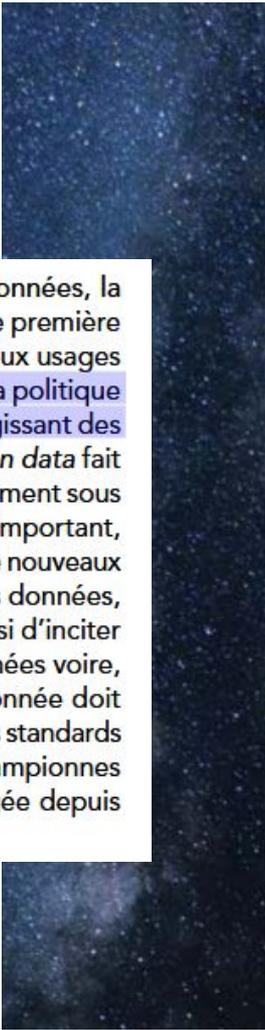
10h15 - Comment annoter des articles ?

10h30 - Echanges

11h15 - Ateliers

12h15 - Restitutions & Prochaines étapes

Pour des IA
francophones :
apprendre aux robots
à parler français



D'abord, une politique offensive visant à favoriser l'accès aux données, la circulation de celles-ci et leur partage. Les données sont la matière première de l'IA contemporaine et d'elles dépend l'émergence de nombreux usages et applications. Il est tout d'abord urgent d'accélérer et d'étoffer la politique d'ouverture des données publiques (*open data*), en particulier s'agissant des données critiques pour les applications en IA. La démarche d'*open data* fait l'objet d'une politique volontariste depuis plusieurs années, notamment sous l'impulsion de la loi pour une République numérique³ : cet effort, important, doit être soutenu. La puissance publique doit par ailleurs amorcer de nouveaux modes de production, de collaboration et de gouvernance sur les données, par la constitution de « *communs de la donnée* »⁴. Il lui revient ainsi d'inciter les acteurs économiques au partage et à la mutualisation de données voire, dans certains cas, d'en imposer l'ouverture. La politique de la donnée doit enfin s'articuler avec un objectif de souveraineté et capitaliser sur les standards de protection européens pour faire de la France et l'Europe les championnes d'une IA éthique et soutenable. L'Union européenne s'est engagée depuis

CÉDRIC VILLANI

Mathématicien et député de l'Essonne

DONNER UN SENS À L'INTELLIGENCE ARTIFICIELLE

POUR UNE STRATÉGIE
NATIONALE ET EUROPÉENNE

Des données d'entraînement de qualité pour des nouveaux usages de l'IA

L'ACOSS : développer un agent vocal conversationnel – voice-bot – pour répondre aux interrogations des utilisateurs du Chèque Emploi Associatif (CEA) ;



Open Justice

Ouvrir la jurisprudence par la pseudonymisation des données



ExploCode

Rendre le droit du travail lisible, accessible et compréhensible

Mettre à disposition des administrations, des laboratoires de recherche, des entreprises, des citoyens des données pour des IA francophones

Une première étape : un jeu de
données ouvert de questions-réponses
francophone

Le projet PIAF : Un produit et des enjeux



usage 1 - Administrations
publiques

usage 2 - Laboratoires de recherche

usage 3 - Entreprises

usage 4 - Citoyens

Place du français
dans l'IA



Micro-travail ?
Contribution engagée ?



Question scientifique :
Natif VS Multilingue



La démarche scientifique

Les modèles de Questions / Réponses

- Sont entraînés pour trouver la "bonne réponse" à une question dans un texte qui contient la réponse. (ex. Dans une biographie de Louis XIV, trouver la réponse "1638" à la question "Date de naissance de Louis XIV").
- Une technologie transformante pour les tâches de recherche ou d'extraction -> plus besoin de structurer les données texte avant des les interroger -> **une nouvelle génération d'IA.**

Le problème :

- Les datasets d'entraînement / évaluation existent exclusivement en anglais / Chinois (SQuAD, QuAC, HotpotQA, NewsQA, etc)
- Peu ou pas de données dans les autres langues. Pas de dataset significatif en Français.
- La traduction automatique des datasets ne suffit pas (env. -10 points de performance selon nos évaluations = 4 ans de retard).

- D'autres pays ont compris l'importance du sujet : SQuAD
Chinois et Coréen

Un protocole inspiré du dataset de référence proposé par Stanford.

The screenshot shows the SQuAD 2.0 website interface. At the top, there is a navigation bar with 'SQuAD', 'Home', 'Explore 2.0', and 'Explore 1.1'. The main header features the title 'SQuAD2.0' and the subtitle 'The Stanford Question Answering Dataset'. Below this, there are two main content areas: 'What is SQuAD?' and 'Leaderboard'. The 'What is SQuAD?' section explains the dataset's purpose and includes a 'New SQuAD2.0' announcement. The 'Leaderboard' section displays a table of top-performing models with their ranks, names, and EM/F1 scores.

What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

New SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 is a challenging natural language understanding task for existing models, and we release SQuAD2.0 to the community as the successor to SQuAD1.1. We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don't know.

[Explore SQuAD2.0 and model predictions](#)

[SQuAD2.0 paper \(Rajpurkar & Jia et al. '18\)](#)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1	ALBERT (ensemble model) Google Language ALBERT Team	89.731	92.215
2	XLNet + DAAF + Verifier (ensemble) PINGAN Omni-Sinitic	88.592	90.859
2	ALBERT (single model) Google Language ALBERT Team	88.107	90.902
2	UPM (ensemble) Anonymous	88.231	90.713
3	XLNet + SG-Net Verifier (ensemble) Shanghai Jiao Tong University & CloudWalk https://arxiv.org/abs/1908.05147	88.174	90.702
4	XLNet + SG-Net Verifier++ (single model)	87.238	90.071

<https://rajpurkar.github.io/SQuAD-explorer/>

Le protocole : annoter des articles Wikipédia en français et disposer de données de qualité et comparables à SQUAD.

Nous avons construit un protocole pour rendre les évaluations comparables avec SQUAD :

- Sélection d'articles similaires en "complexité"
- Protocole d'annotation différent (pas de Mechanical Turk) mais comparable en output.

Les données PIAF permettront :

phase 1 (collecte de données d'évaluation):

mesurer avec fiabilité les écarts de performance des différents modèles multilingues existants.

phase 2 (collecte de données d'entraînement):

pour entraîner nativement des modèles monolingues français ou adapter des modèles multilingues au français non-traduit

> Une opportunité concrète d'amélioration des IA francophones <

Les différentes étapes

1

Collecte des données d'évaluation & de test

20% des articles sources
Annotations "certifiées"



2

Premiers tests



3

Collecte des données d'entraînement

80% des articles sources
Annotation "grand public"

<i>mode</i>	<i>paire QR/minute</i>	<i>collaborateur/h pour 10k paires</i>	<i>paires par session</i>
base (entraînement)	1.25	133.3	6000
base (additionel/dev)	2.5	66.6 (x n_{req})	12000/ n_{req}
questions impossibles	1.4	233.3	6720

Table 1: Temps estimé pour chaque mode. n_{req} est le nombre minimal de réponses additionnelles nécessaires (pour SQuADv1.1, $n_{req} = 2$)

Comment annoter des
articles ?



1

2

3

4

5

À l'automne 1898 sort, imprimé en Grasset, Les Aventures merveilleuses de Huon de Bordeaux, chanson de geste (car le Moyen Âge est l'époque favorite de l'Art nouveau). Le monde de la typographie est alerté et paraît favorable. En 1900, sept corps seulement sont gravés ; les demandes affluent et il faut commercialiser. Georges Peignot et Francis Thibaudeau, maître typographe de grande qualité qu'il vient d'engager, créent une petite plaquette d'un discret mais très bon goût. A l'envoi des plaquettes correspond un afflux de commandes (auquel contribuent les premiers caractères Auriol, cf. ci-dessous). Parallèlement, les éloges de la presse spécialisée et des connaisseurs d'art déferlent. Dans les cours d'immeubles du boulevard de Montrouge (plus tard rebaptisé en Edgar-Quinet) où ils sont installés depuis 34 ans, les ateliers Peignot sont brusquement engorgés.

Écrire une question

OK

Une démarche contributive et
apprenante

Un projet ouvert,
documenté, qui fait le
choix de la "contribution
engagée"

Une plateforme d'annotation *open source*

Un premier jeu de données *open data*
de questions-réponses construit selon
une méthodologie scientifique

Une *méthode ouverte* : contributions
volontaires et communauté

Parler

Donnez un peu de votre voix



Écouter

Aidez-nous à valider les
échantillons vocaux



Un projet partenarial

reciTAL.

Accompagnement scientifique

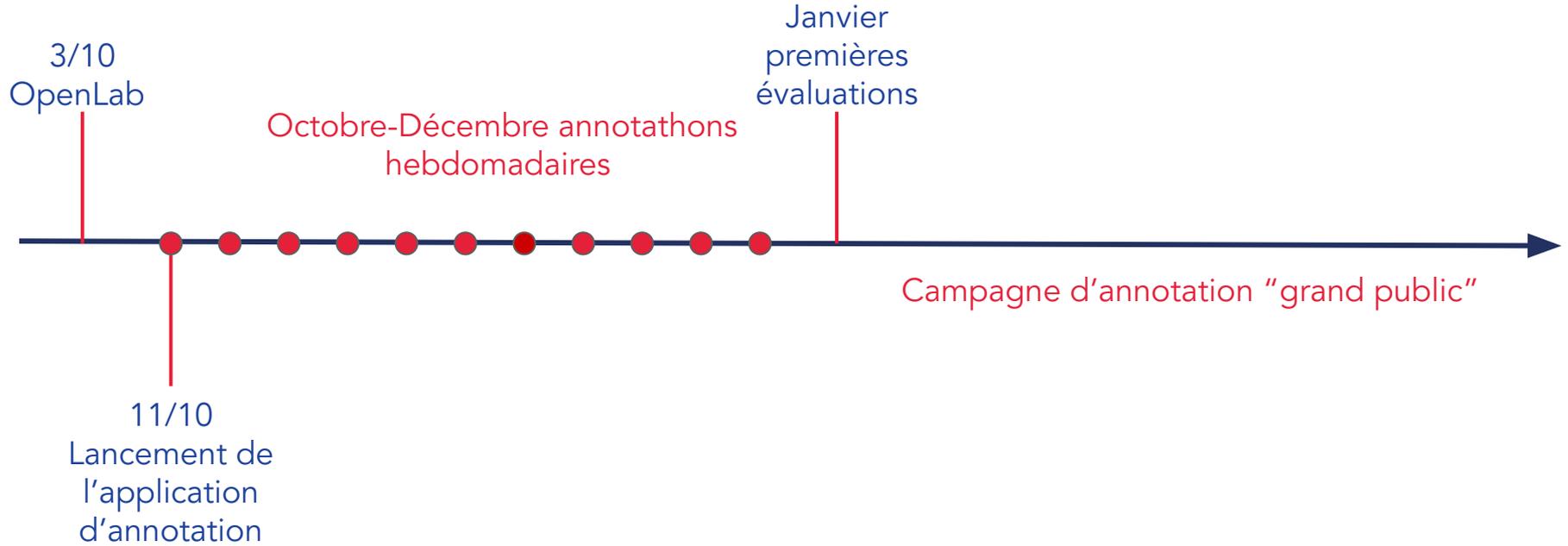


Accueil des événements d'annotation hebdomadaires



Participation au financement du projet via le Programme d'investissements d'avenir

Calendrier prévisionnel



Événements d'annotation hebdomadaires

A partir du 11 octobre - RDV tous les
vendredis de 12h30 à 14h au 77 avenue
de Ségur

Pour faire partie des inscrits :

<https://listes.etalab.gouv.fr/listinfo/piaf>

Contribuer au projet PIAF

- Organiser un événement d'annotation avec votre communauté
- Proposer des cas d'usages de données de questions-réponses en français : pour la recherche, l'action publique, etc.
- Partager de l'expérience sur les initiatives de sciences participatives ou des projets de crowdsourcing

> piaf@data.gouv.fr <

Echanges

Questions-Réponses

Ateliers

4 ateliers - format world café

Tester PIAF et apporter des retours utilisateurs - Guillaume

Explorer des cas d'usages de données de questions-réponses en français -
Paul-Antoine

Comment valoriser l'engagement des annotateurs ? - Mathilde

Quels enjeux scientifiques autour du projet PIAF ? - équipe reciTAL

Restitutions des ateliers

4 ateliers - format world café

Tester PIAF et apporter des retours utilisateurs - Guillaume

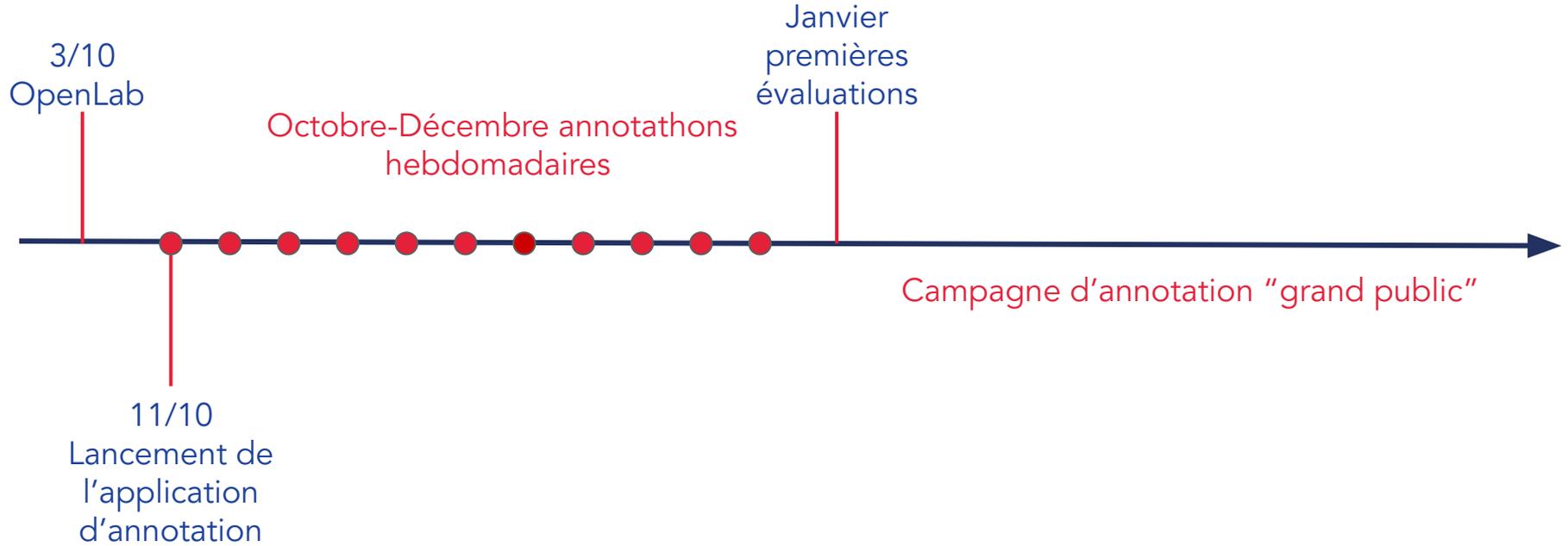
Explorer des cas d'usages de données de questions-réponses en français -
Paul-Antoine

Comment valoriser l'engagement des annotateurs ? - Mathilde

Quels enjeux scientifiques autour du projet PIAF ? - équipe reciTAL

Prochaines étapes

Calendrier prévisionnel



Événements d'annotation hebdomadaires

A partir du 11 octobre - RDV tous les
vendredis de 12h30 à 14h au 77 avenue
de Ségur

Pour faire partie des inscrits :

<https://listes.etalab.gouv.fr/listinfo/piaf>

Contribuer au projet PIAF

- Organiser un événement d'annotation avec votre communauté
- Proposer des cas d'usages de données de questions-réponses en français : pour la recherche, l'action publique, etc.
- Partager de l'expérience sur les initiatives de sciences participatives ou des projets de crowdsourcing

> piaf@data.gouv.fr <

A bientôt !
sur piaf.etalab.studio