

# Pour des IA Francophones

Open Lab n°2 - 26 février 2020

Lieu de la transformation publique

[piaf.etalab.studio](https://piaf.etalab.studio)



etalab gouv.fr

# Programme

14h - Accueil

14h30 - Restitutions - PIAF, 6 mois après ? // *Plénière*

- Enseignements scientifiques
- Retours sur la méthode contributive
- Prochaines étapes et échanges

15h30 - Ateliers // *Espace événementiel*

- Identifier les cas d'usage et faire vivre la communauté (Mathilde, Benjamin, Guillaume)
- Réutiliser les données de PIAF (Julien, Kim)
- Les futurs PIAF du Lab IA (Paul-Antoine)

17h - Conclusion

PIAF - créer le 1<sup>er</sup> jeu de données ouvert  
de questions-réponses en français  
pour mieux entraîner les IA

# PIAF, un projet d'outil mutualisé du Lab IA



Des projets d'IA  
publics accompagnés  
via des appels à  
manifestation  
d'intérêts



Des outils  
mutualisés : PIAF,  
pseudonymisation



Une communauté :  
data scientists,  
chercheurs...

# Contribuer à la souveraineté de l'IA francophone, en particulier dans le champ du traitement automatique du langage

CÉDRIC VILLANI

Mathématicien et député de l'Essonne

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 numérique<sup>3</sup> : 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* »<sup>4</sup>. 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

## DONNER UN SENS À L'INTELLIGENCE ARTIFICIELLE

POUR UNE STRATÉGIE  
NATIONALE ET EUROPÉENNE

# Explorer des cas d'usage dans le champ de l'IA publique

The screenshot shows the website for Rosny-sous-Bois. At the top, it says "Le magazine de la ville" with a button "VOIR TOUS LES MAGAZINES". Below that, "Recevez nos informations" with a button "M'INSCRIRE À UNE NEWSLETTER". The main content area displays statistics: 45663 HABITANTS and 14810 EMPLOIS. A chatbot window is open, displaying a greeting: "Bonjour Je m'appelle Rosnyz, je suis là pour répondre à toutes vos questions." Below the greeting are buttons for "Nous rejoindre", "Mes démarches", "Portail familles", "Menus des cantines", and "Gestion des déchets". At the bottom of the chatbot, there is a question: "Comment faire ma carte d'identité ?" and a response: "Désolé, je ne comprends pas encore votre question mais j'apprends tous les jours !". The website footer includes contact information, social media links, and a copyright notice.

The screenshot shows the website for "Code du travail numérique". At the top, it features the French flag and the text "Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE". The main heading is "Code du travail numérique" with a "Version Bêta" badge. Below the heading, it says "Bienvenue sur le Code du travail numérique". A sub-heading reads: "Vous cherchez une information sur le droit du travail ? Vous avez besoin d'accompagnement ? Nous vous proposons des réponses accessibles et personnalisées selon votre situation." There is a search bar with the placeholder "Recherche" and a "Rechercher" button. Below the search bar, it says "Recherchez par mots clefs". A section titled "Boîte à outils" with a right arrow contains the text "Trouvez des réponses personnalisées selon votre situation". At the bottom, there are three cards: "Simulateur d'indemnités de licenciements" with a "Démarrer" button, "Simulateur de durée de préavis de démission" with a "Démarrer" button, and "Modèles de courriers" with a "Consulter" button.

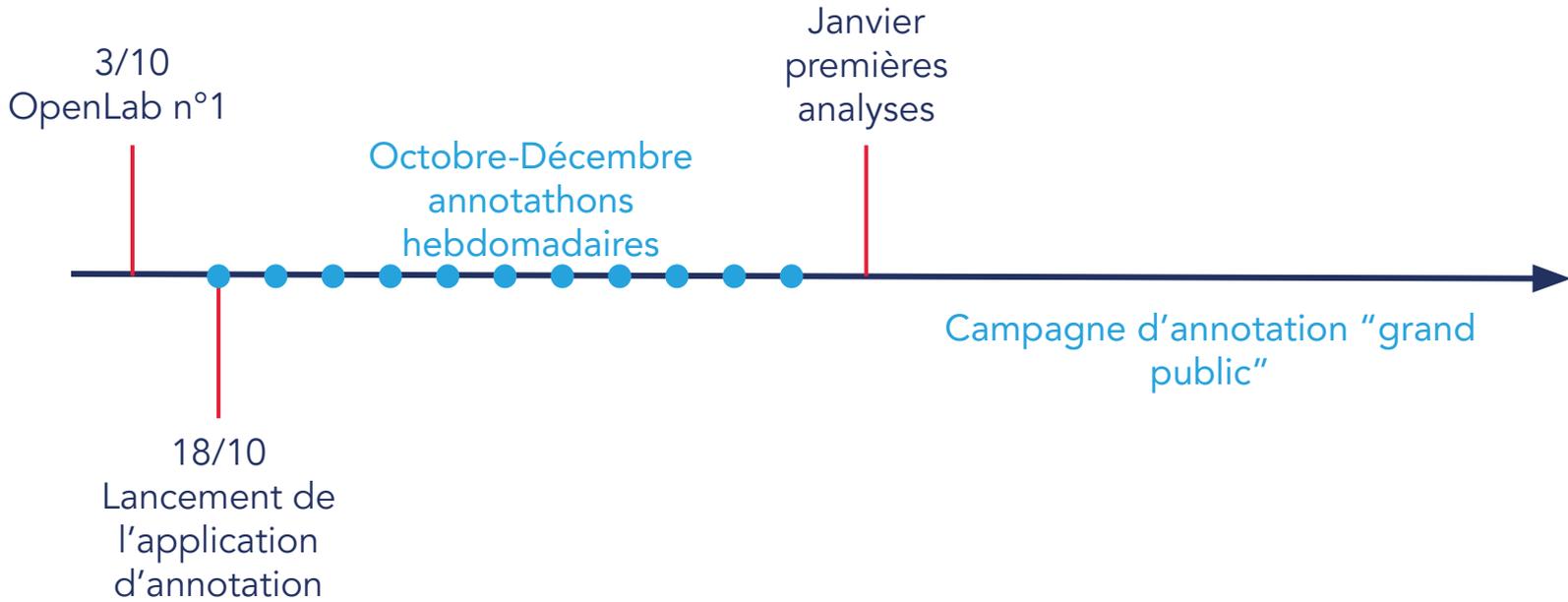
Mettre en œuvre une démarche ouverte et contributive autour de l'IA publique



# Les 3 questions exploratoires de PIAF

1. Est-il nécessaire de disposer d'un jeu de données de questions-réponses nativement en français ?
2. Comment la communauté de contributeurs a-t-elle permis de faire évoluer le projet ?
3. Quels sont les enseignements à tirer d'une démarche d'annotation volontaire : quelles limites, quelles opportunités ?

# Pour répondre à ces questions, un calendrier ambitieux

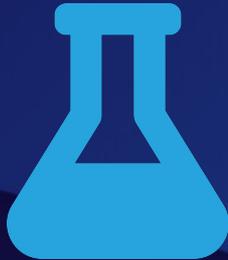


# 6 mois plus tard, PIAF c'est...



Une plateforme  
d'annotation  
open Source

<https://github.com/etalab/piaf>



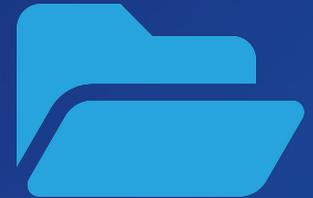
Des analyses  
scientifiques  
enrichies

<https://piaf.etalab.studio/protocole-fr/>



Une méthodologie  
contributive  
expérimentée

<https://piaf.etalab.studio/enseignements-contributions/>



Un premier jeu de  
données

qui sera ouvert  
pour la  
conférence [LREC](#)  
en mai 2020

Est-il nécessaire de disposer d'un jeu de données de questions-réponses en français ?

# Un double problème

Question scientifique initiale : quel est le gain de performance généré par l'utilisation d'un jeu de données d'apprentissage francophone ?

- par rapport à un modèle entraîné uniquement sur de l'anglais
- par rapport à un modèle entraîné sur des données traduites automatiquement en français

Pour répondre à cette question, il faut d'abord mesurer les performances du modèle sur des données d'évaluation de QR en français... or ces données n'existent pas !

# Comparaisons des données d'évaluation

Les données collectées par PIAF nous permettent de répondre en partie à ces questions :

<i>train</i> ↓ <i>dev</i> →	EN	FR_ <i>(trans)</i>	FR_ <i>PIAF</i>
EN	81.88	70.64	61.98
FR_ <i>trans</i>	81.72	75.13	64.02
EN+FR_ <i>trans</i>	81.74	75.52	65.60

Les scores de modèles multilingues sur le français traduit et le français ne sont pas les mêmes : le français natif est plus difficile. Il semble subsister des biais de traduction.

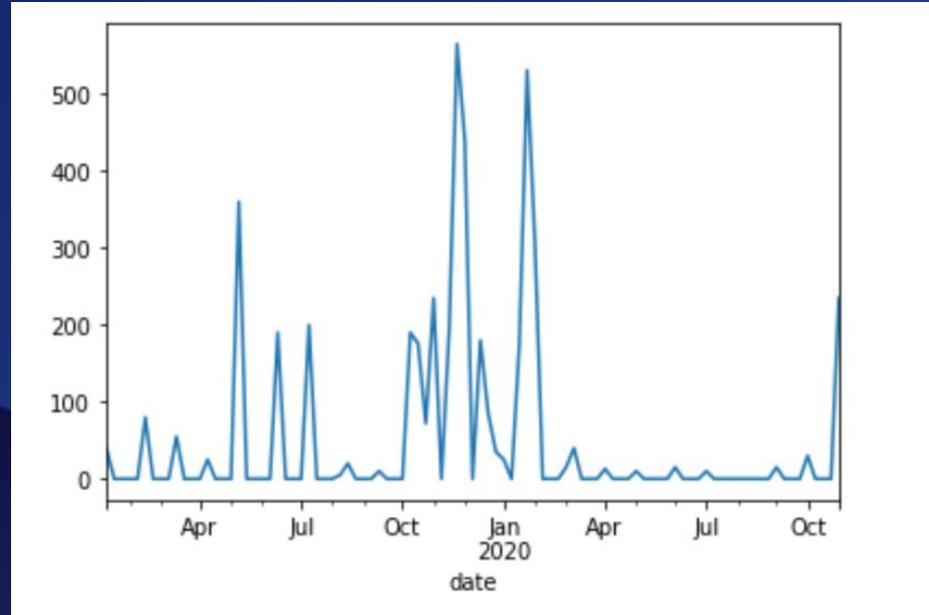
# Le jeu de données PIAF

Des premières expérimentations qui nous amènent à redimensionner nos objectifs :

- de 20k à 5k pour la validation
- de 80k à 20k pour l'entraînement

Nombre d'exemples d'entraînement	22263	43292	65 555	86584
F1 score	85.70	87.83	88.53	89.17

*Performance en fonction de la taille du dataset d'entraînement*



*Nombre de contributions au jeu de données de QR par semaine*

# CamemBERT, FlauBERT, PIAF : trois projets d'IA francophones

Modèles de langage francophones



Dataset QA francophone



**Extrait du Code de Travail**

Le contrat de travail à durée indéterminée (CDI) est la forme normale et générale de la relation de travail. Par définition, **il ne prévoit pas la date à laquelle il prend fin**. Il peut être rompu sur décision unilatérale soit de l'employeur (licenciement pour motif personnel ou pour motif économique), soit du salarié (démission, départ à la retraite).

**Question**

Quelle est la durée d'un CDI ?

**Réponse proposée par le modèle**

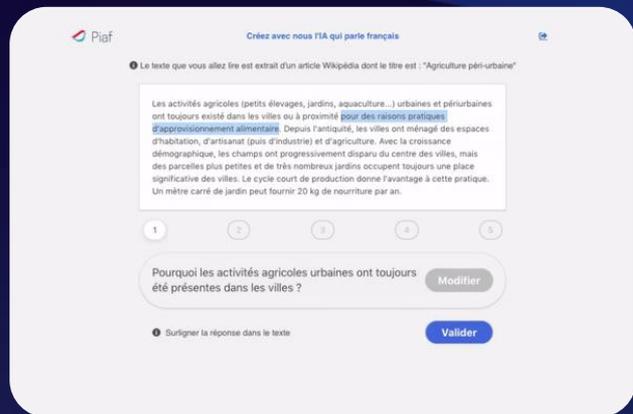
**il ne prévoit pas la date à laquelle il prend fin**

*(Un éventuel modèle FR-BERT fine-tuné sur PIAF)*

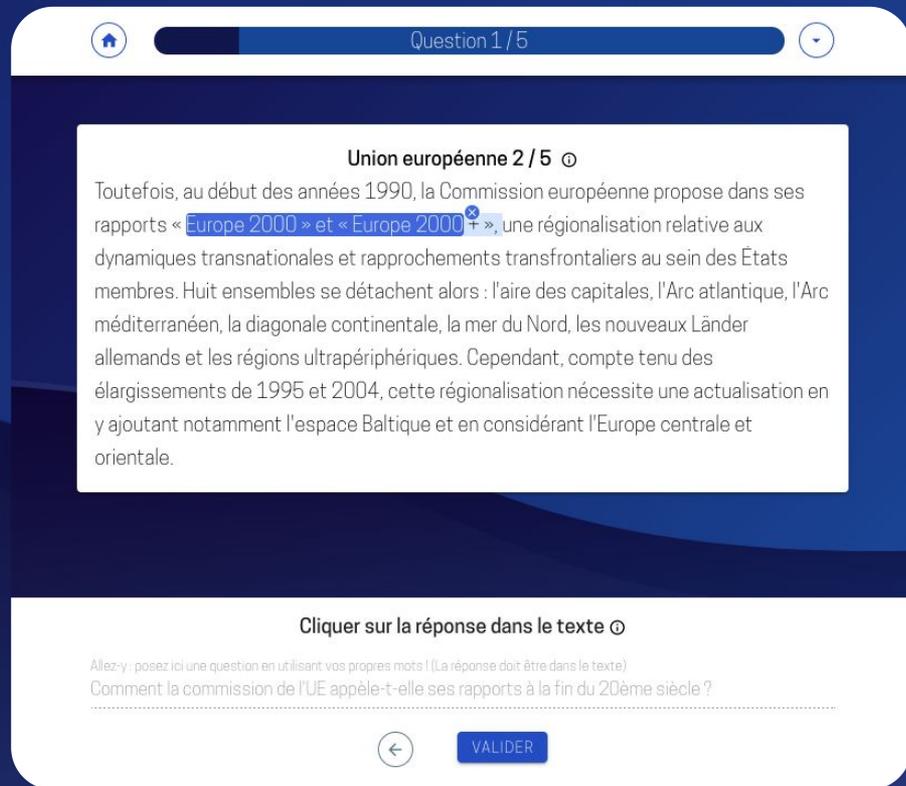
<https://piaf.etalab.studio/francophonie-ia/>

Comment la communauté de contributeurs a-t-elle fait évoluer la plateforme PIAF ?

# Co-construction de la plateforme d'annotation



The screenshot shows the Piaf annotation tool interface. At the top, it says "Créez avec nous l'IA qui parle français". Below that, a text box contains a paragraph from Wikipedia about urban agriculture. A question is displayed below the text: "Pourquoi les activités agricoles urbaines ont toujours été présentes dans les villes ?". There are navigation buttons (1-5) and a "Modifier" button. At the bottom, there is a "Valider" button and a note: "Surigner la réponse dans le texte".



The screenshot shows a question card from the annotation platform. At the top, it says "Question 1 / 5". The main text is: "Union européenne 2 / 5". Below that, the question text is: "Toutefois, au début des années 1990, la Commission européenne propose dans ses rapports « Europe 2000 » et « Europe 2000+ », une régionalisation relative aux dynamiques transnationales et rapprochements transfrontaliers au sein des États membres. Huit ensembles se détachent alors : l'aire des capitales, l'Arc atlantique, l'Arc méditerranéen, la diagonale continentale, la mer du Nord, les nouveaux Länder allemands et les régions ultrapériphériques. Cependant, compte tenu des élargissements de 1995 et 2004, cette régionalisation nécessite une actualisation en y ajoutant notamment l'espace Baltique et en considérant l'Europe centrale et orientale." Below the text, there is a "Valider" button and a note: "Cliquez sur la réponse dans le texte". At the bottom, there is a "VALIDER" button and a note: "Allez-y, posez ici une question en utilisant vos propres mots ! (La réponse doit être dans le texte)".

Quels sont les enseignements à tirer  
d'une démarche de contribution  
volontaire ?

# Les enjeux d'une démarche contributive : la communauté et l'acceptabilité

<https://voice.mozilla.org/fr>

Common Voice  
mozilla

CONTRIBUER JEUX DE DONNÉES LANGUES QUI SOMMES-NOUS  0  0

Se connecter  
/ S'inscrire

FR ▼

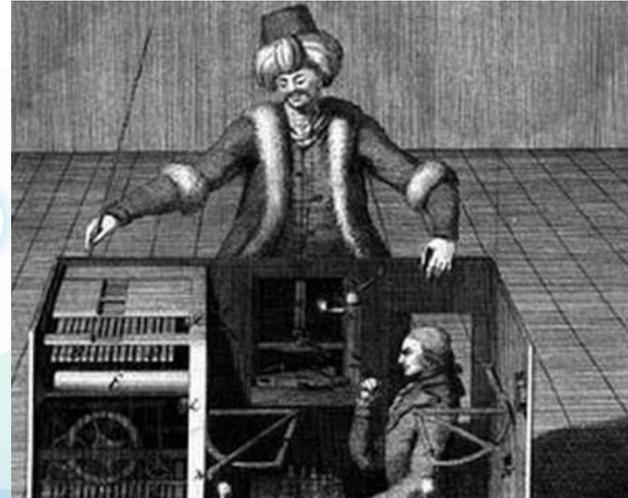
## Parler

Donnez un peu de votre voix



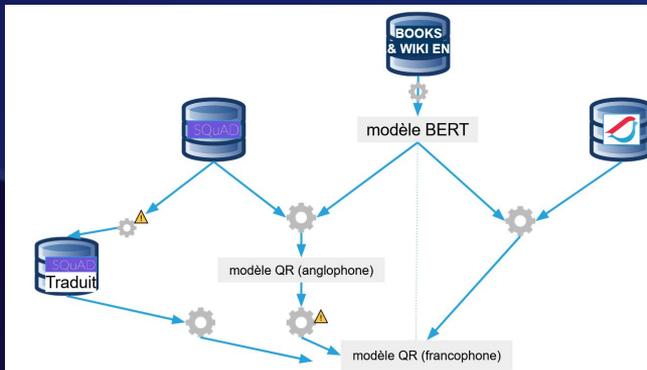
## Écouter

Aidez-nous à valider les échantillons vocaux



# Une méthodologie contributive traduisant les objectifs de PIAF

## Pédagogie



## Exploration scientifique

**François Chollet** @fchollet

One of the most effective things a government can do to foster an ecosystem of AI startups is to have a National Open Data Strategy.

Put all government data behind APIs, sponsor the creation of high-quality public datasets in various domains relevant to local markets & language.

8:51 PM · Oct 14, 2019 · Twitter for Android

406 Retweets 1.1K Likes

## Ouverture



12

Annotathons

10

Tournées

<https://piaf.etalab.studio/actualites.html>

# La communauté PIAF

350 contributeurs



reciTAL.



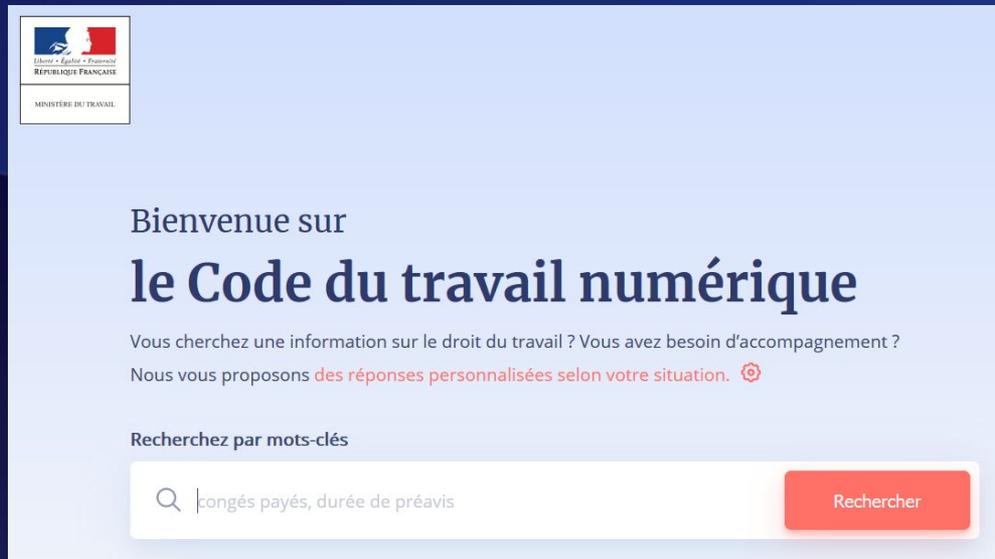
# Les frontières de la contribution volontaire : quelques bonnes pratiques

-  Échanger collectivement sur les enjeux éthiques, quitte à co-construire une charte de la contribution
-  Tester la méthode de contribution sur des périodes courtes pour l'adapter
-  Mobiliser une équipe à plein temps sur l'animation de la communauté
-  Documenter le projet en continu

Pour approfondir : <https://piaf.etalab.studio/enseignements-contributions/>

Prochaines étapes ?

# Tester les premières données sur des usages : le cas d'Expoclode / Code du travail numérique



The screenshot shows the official website of the French Ministry of Labor. At the top left is the logo of the French Republic and the text 'MINISTÈRE DU TRAVAIL'. The main heading reads 'Bienvenue sur le Code du travail numérique'. Below this, a question asks if the user is looking for information on labor law or needs assistance, followed by a promise of personalized answers. A search bar is present with the text 'Recherchez par mots-clés' and a search button labeled 'Rechercher'. The search bar contains the text 'congés payés, durée de préavis'.

MINISTÈRE DU TRAVAIL

Bienvenue sur  
**le Code du travail numérique**

Vous cherchez une information sur le droit du travail ? Vous avez besoin d'accompagnement ?  
Nous vous proposons **des réponses personnalisées selon votre situation.**

Recherchez par mots-clés

congés payés, durée de préavis

Rechercher

RDV à l'atelier Réutiliser les données de PIAF (Julien/Kim)

# Identifier des cas d'usage métier et faire vivre la communauté



RDV à l'atelier : Identifier des cas d'usage et faire vivre la communauté (Benjamin/Mathilde/ Guillaume)

# Identifier de nouveaux projets d'IA francophones (champ du TAL ou autres)



RDV à l'atelier : Les futurs PIAF du Lab IA (Paul-Antoine)

# Questions-réponses

Ateliers

# Cas d'usage métiers de PIAF

Guillaume/Benjamin/Mathilde

# Réutiliser les données de PIAF

Julien

# Les futurs PIAF du Lab IA

Paul-Antoine

Merci !

[piaf@data.gouv.fr](mailto:piaf@data.gouv.fr)

[piaf.etalab.studio](http://piaf.etalab.studio)

