



# medkit - une bibliothèque python pour un système de santé apprenant

Olivier Birot, Équipe HeKA



**Inserm**



Université  
Paris Cité

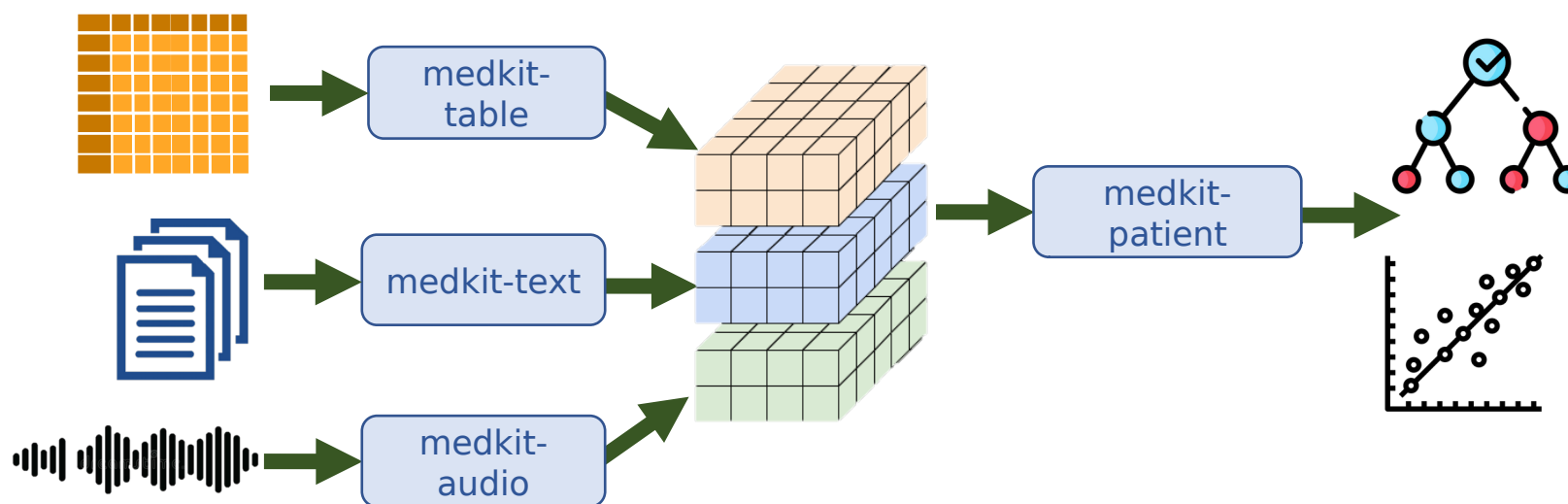
Rencontres ingénieurs INRIA en santé numérique  
7 décembre 2022

# Objectif général

Système de santé produit données de différentes modalités (imagerie médicale, compte rendus rédigés, résultats d'analyses, etc).

Medkit veut faciliter :

1. l'**extraction d'informations** depuis des données de santé brutes de diverses modalités (*medkit-text*, *medkit-audio*, etc, en cours de développement)
2. le développement de **modèles multi-modaux** fournissant une aide à la décision médicale (*medkit-patient*, développement à venir)



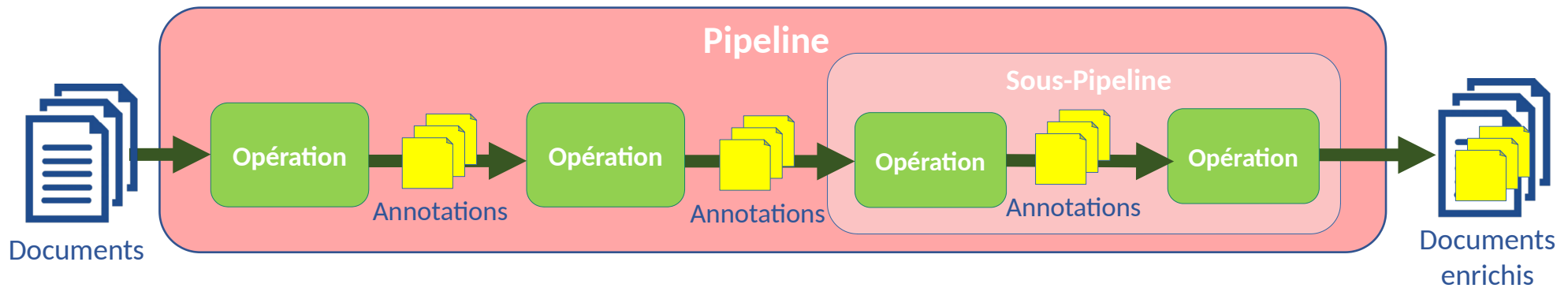
# Utilisateurs cibles

- utilisateur « simple » consommant les composants disponibles dans medkit (en les reparamétrant si nécessaire)
- utilisateur contribuant de nouveaux composants à medkit, qui peuvent s'appuyer sur les composants existants

Objectif : faire de medkit un **espace de partage** de nouveaux modèles



# Architecture



- Structures de données : **documents** (associés à une modalité) auxquels on attache des **annotations** (inspiré de bibliothèques de NLP)
- **Opérations** : reçoivent annotations, produisent nouvelles annotations
- **Pipelines** : graphes d'opérations, composables en sous-pipelines. Permet d'obtenir un graphe de provenance avec structure similaire.

Modalités actuellement supportées : principalement texte, un peu d'audio

# Architecture - Opérations

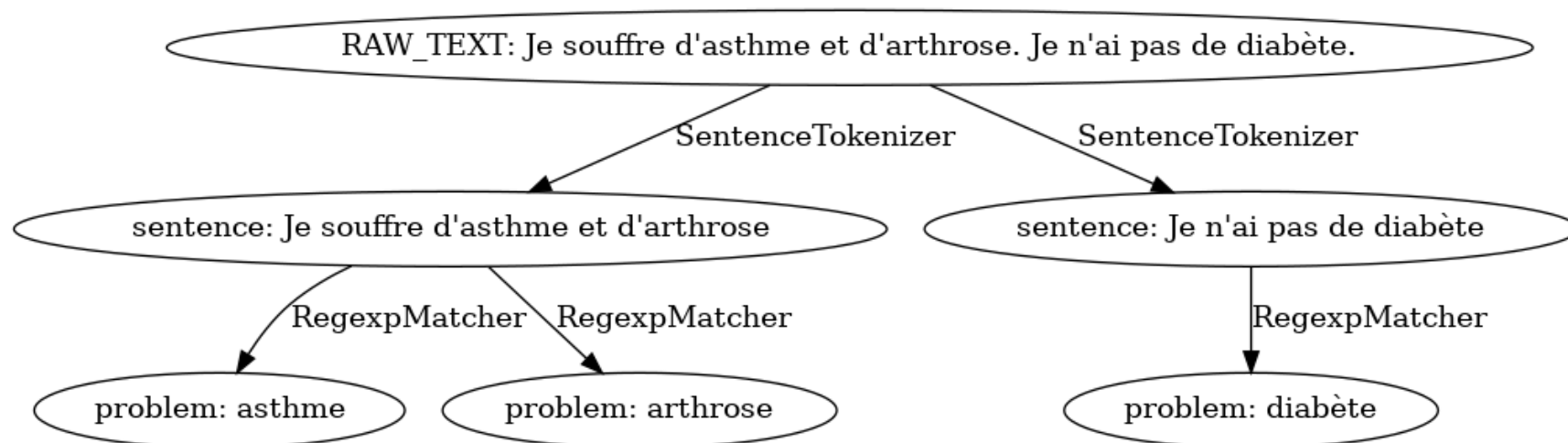
- Opérations medkit « **core** » simples, configurables, peu de dépendances
  - texte : tokenization, détection d'entités via regexps, détection de négation, etc
  - audio : normalization, ré-échantillonnage, etc
- Opérations plus complexes basées sur **composants tierce partie**, dépendances optionnelles
  - texte : détection d'entités médicales avec modèles HuggingFace, détection de relations avec spaCy, etc
  - audio : détection de locuteur avec pyannote-audio, transcription avec speechbrain
- **Extensible** : utilisateur développe ses propres opérations

Medkit = **boîte à outils** permettant d'interfacer différents composants



# Architecture - Pipeline et provenance

- **Graphe de provenance** : medkit mémorise pour chaque nouvelle annotation l'opération qui l'a produite et les annotations sources utilisées
- Si pipeline décomposée en sous-pipelines, alors graphe de provenance décomposé en sous-graphes selon la même structure. Granularité ajustable.
- A venir : export au format W3C PROV-O/DM



# Etat du projet

- projet démarré en janvier 2022, pas encore mature, développé par 2 ingés HeKA + 1 ingé SED
- Sources : <https://gitlab.inria.fr/heka/medkit> et <https://github.com/TeamHeka/medkit>, documentation : <https://heka.gitlabpages.inria.fr/medkit/>
- Contributions bienvenues, en particulier intégration de composants tierce-partie dans opérations. Licence MIT.

## Difficultés et enjeux à venir:

- diversité des use cases, hétérogénéité des données (formats, structures)
- scalabilité sur volumes de données plus élevés pas encore étudiée
- fine-tuning pour opérations utilisant modèles appris
- *medkit-patient* ! (modèles multi-modaux)

