

# Preparing the Dictionnaire Universel for Automatic Enrichment

---

Pedro Javier ORTIZ SUAREZ<sup>1,2</sup>, Laurent ROMARY<sup>1</sup> Benoît SAGOT<sup>1</sup>

June 12, 2019

<sup>1</sup>Inria, Paris, France

<sup>2</sup>Sorbonne Université, Paris, France

# Table of contents

1. Introduction
2. Automatic Enrichment of Ancient Dictionaries
3. Workflow
4. Preparing the Dictionnaire Universel for Automatic Enrichment

# Introduction

---

# The Dictionnaire Universel

- In 1635, the *Académie française* began producing a dictionary of the French language [2]. *Antoine Furetière* got involved in the production of the first edition.
- Furetière grew frustrated with the slow progress of the *Académie*. He began to work on his own dictionary, *The Dictionnaire universel (DU)* around 1676-78 [7].
- The second edition (1701) of the *DU* was improved by *Henri Basnage de Beauval* who virtually turned the *DU* into an encyclopaedia [1][3].





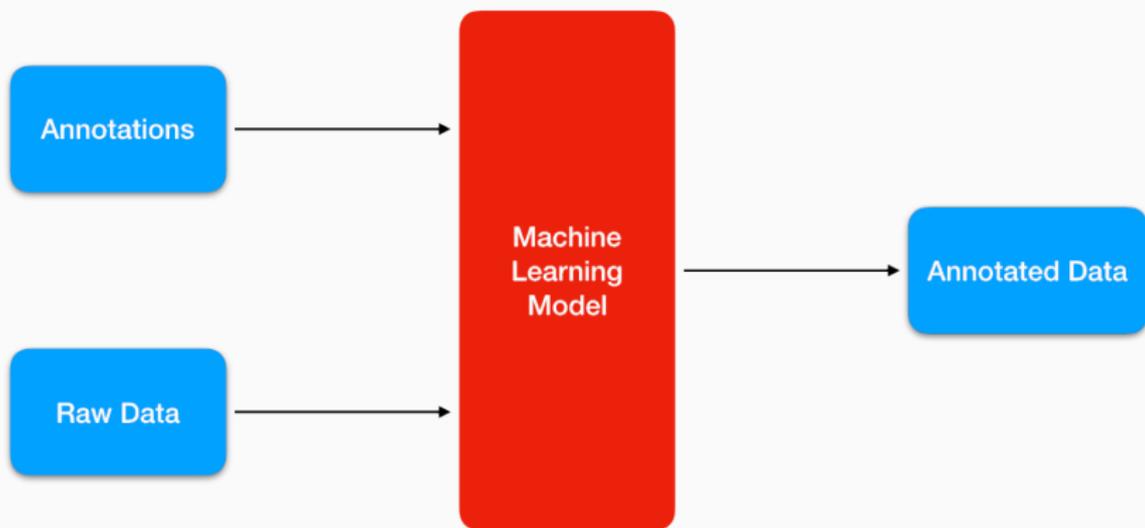
# Automatic Enrichment of Ancient Dictionaries

---

At *Inria* we will develop methods for the automatic structuring of lexicographical entries from the digitized version of Basnage's *DU*, we will also enrich the contents of this dictionary using **machine learning** methods. There are two main types of machine learning models:

- **Statistical methods:** We have some existing tools.
- **Neural methods:** We want to build our own tools.

# Machine Learning



# Optical Character Recognition

## Transkribus

- Well established tool.
- Uses **statistical methods**.
- Has an user interface.
- Generates PDF's.
- Does not preserve typographical features.
- Free but not Open Source.
- Transkribus infrastructure has to be used.

## Kraken

- Not so well established.
- Uses **neural methods**.
- User interface in development.
- Does not generate PDF's.
- Preserves Typographical features.
- Free and Open Source.
- We can use our own infrastructure.

GROBID-Dictionaries [4][5] is an machine learning library using **statistical methods** for structuring digitised lexical resources and entry-based documents with encyclopedic or bibliographic content. It allows the parsing, extraction and structuring of text information in such resources.

GROBID-Dictionaries takes as input lexical resources digitised in PDF format and generates a TEI-encoded hierarchy of the different recognised text structures.

**RAQUETTE**. *s. f.* Espece de palette pour jouer à la paume, & au volant. Elle est faite d'un treillis de cordes de boyaux (dont les unes s'appellent *montans*, & les autres *travers*) fort tendues sur un tour de bois qui a un manche de mediocre longueur. Un de ses côtez s'appelle les *droits*, & l'autre les *nœuds*. Pasquier a remarqué qu'anciennement on ne jouoit point à la paume avec des *raquettes*: c'étoit avec la paume de la main; & de là il conjecture qu'est venu le nom de jeu de *paume*. On n'avoit inventé les *raquettes* qu'un peu avant le temps de Pasquier, à ce qu'il dit.

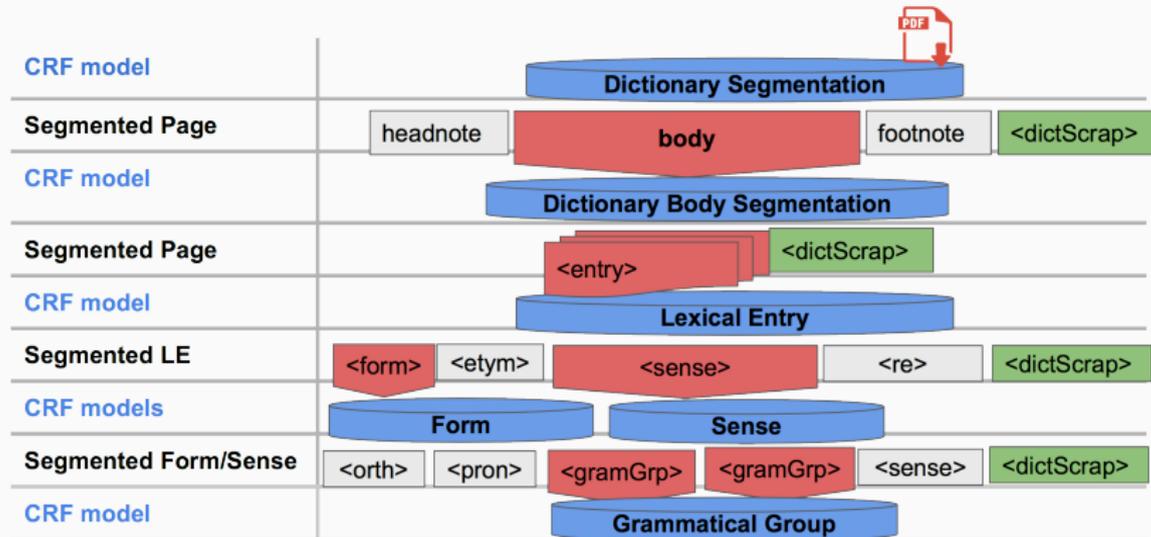
Menage derive ce mot de *resiquetta*, diminutif de *retis*, *reticus* & *reticulum*.

On dit proverbialement pour se moquer d'un homme qui se vante de plusieurs choses qu'il n'a pas faites, C'est un grand cañeur de *raquettes*.

**RAQUETTE**, se dit aussi d'une certaine machine que les Sauvages de Canada attachent à leurs pieds pour marcher plus commodément sur la neige, & qui est faite à-peu-près en forme de *raquette* à jouer.

**RAQUETTE**, se dit aussi d'une espece de figuier d'Inde qui croît aux Iles Antilles: c'est cette espece que Mr. Tournefort appelle *opuntia vulgò herbariorum*, J. BAUH. Ses feuilles sont épaisses, longues, quelquefois larges comme une raquette, d'où vient que les François lui ont donné ce nom. Voyez FIGUIER D'INDE.

# GROBID-Dictionaries



```
<div n="RAQ">
  <entry xml:lang="fre" xml:id="Raquette">
    <sense<form<orth rendition="#uc">raquette.</orth><gramGrp><pos expand="Substantif">s.</pos>
    <gen expand="Feminin">f.</gen></gramGrp></form> <def>Espece de pallette pour jouer
    à la paume, &amp; au volant.</def> <note>Elle est faite d'iun treillis de
    cordes de boyaux édont les unes s'appellent montans
    &amp; les autres travers J fort tendues sur un tour de bois
    qui a un manche de mediocre longueur.</note> <xr corresp="#droits #noeuds">Un de ses cô-
    tea s'appelle les droits, &amp; l'autre les nœuds</xr>. <etym<bibl><author ref="#Pasquier_ISN000000010907010">Pasquier</author></bibl>
    a remarqué qu'anciennement on ne joûoit point à la
    paume avec des raquettes : c'étoit avec la paume de la
    main , &amp; de là il conjectue qu'est venu le nom de jeu
    de pame.</etym> <note>On n'avoit inventé les raquettes qu'un peu
    avant le temps de <bibl><author ref="#Pasquier_ISN0000000010907010">Pasquier</author></bibl>, à ce qu'il dit.</note>

    <etym<bibl><author ref="#Men_ISN00000000080815971">Menage</author></bibl> derive ce mot de
    <lang rendition="#i" xml:lang="lat">retiquetta</lang>, diminutif de retis,
    reticus &amp; reticulum.</etym>
  </sense>

  <sense<usg>On dit proverbialement</usg> <gloss>pour se moquer d'un homme qui
  se vante de plusieurs choses qu'il n'a pas faites,</gloss> <seg type="Proverbe">C'est
  un grand casseur de raquettes.</seg></sense>

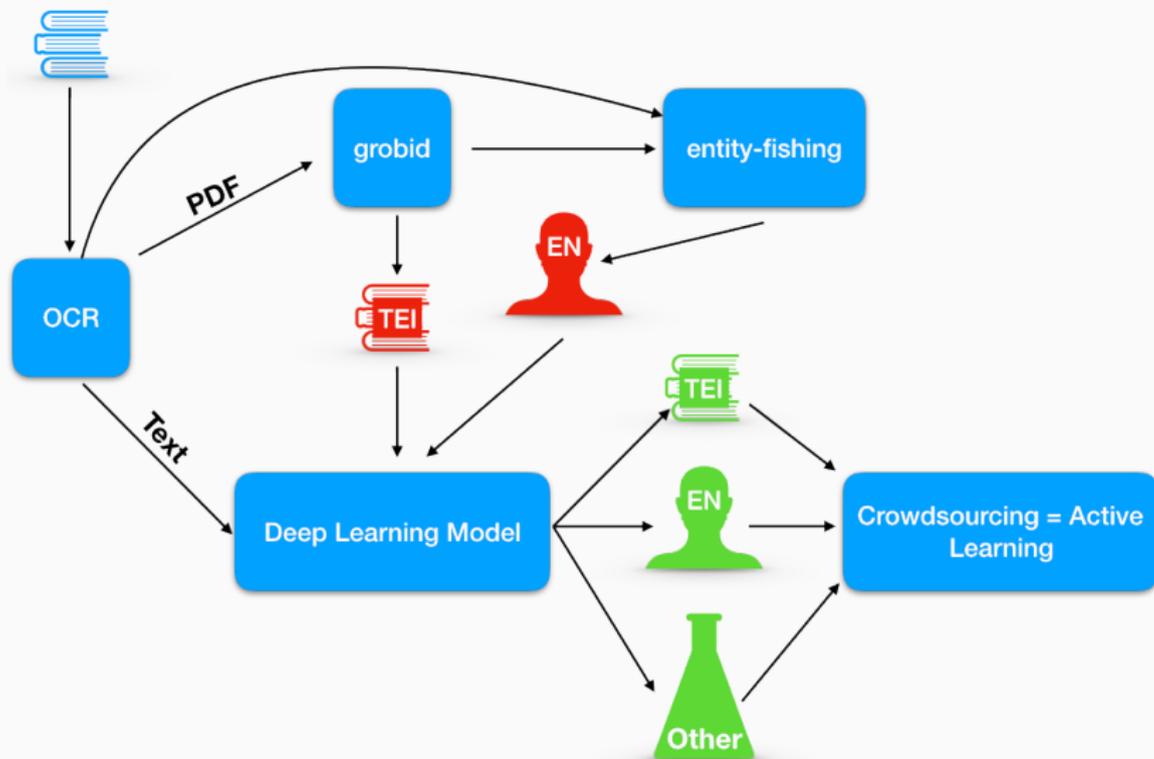
  <sense<form<orth rendition="#sc">Raquette,</orth></form> <def>se dit aussi d'une certaine machine que
  les Sauvages de Canada attachent à leurs pieds pour
  marcher plus commodément sur la neige, &amp; qui est
  faite à-peu-près en forme de raquette à jouer.</def></sense>

  <sense<form<orth rendition="#sc">Raquette,</orth></form> <def>se dit aussi d'une espece de figuier d'In-
  de qui croit aux lles Antilles :</def> <note>c'est cette espece que
  Mr. Tournefort appelle opuntia vule herbariorum,
  J. Bauh. Ses feuilles sont épaisses, longues, quel-
  quefois larges comme une raquette, d'où vient que les
  François lui ont donné ce nom.</note> <xr corresp="#Ficuier">Ficuier
  d'Inde.</ref></xr></sense>
  </entry>
</div>
```

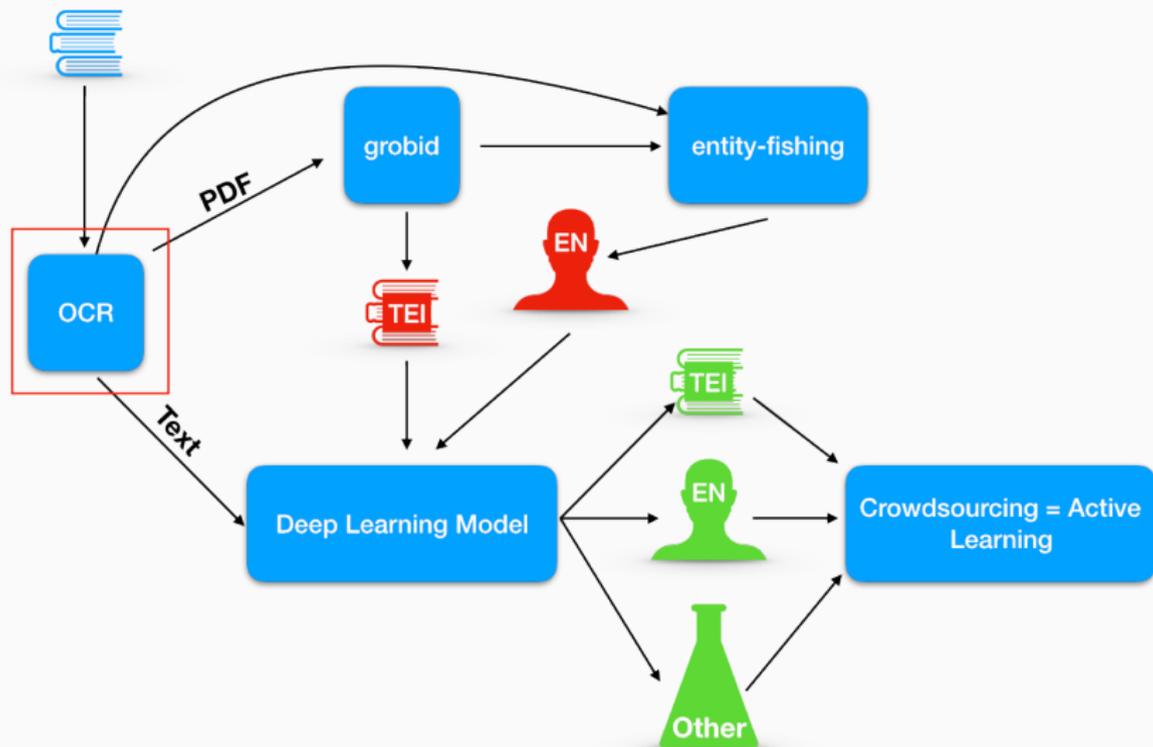
# Workflow

---

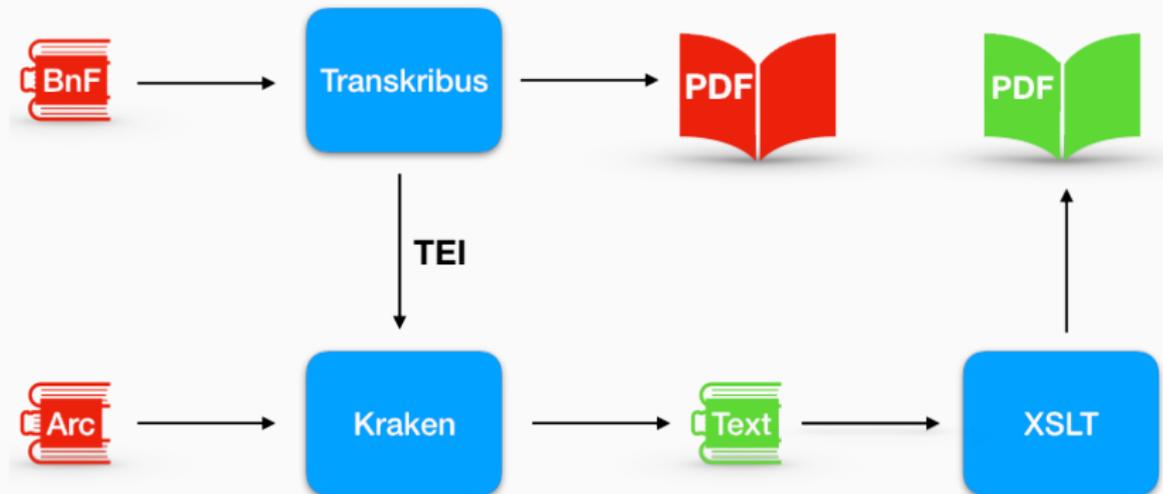
# Workflow



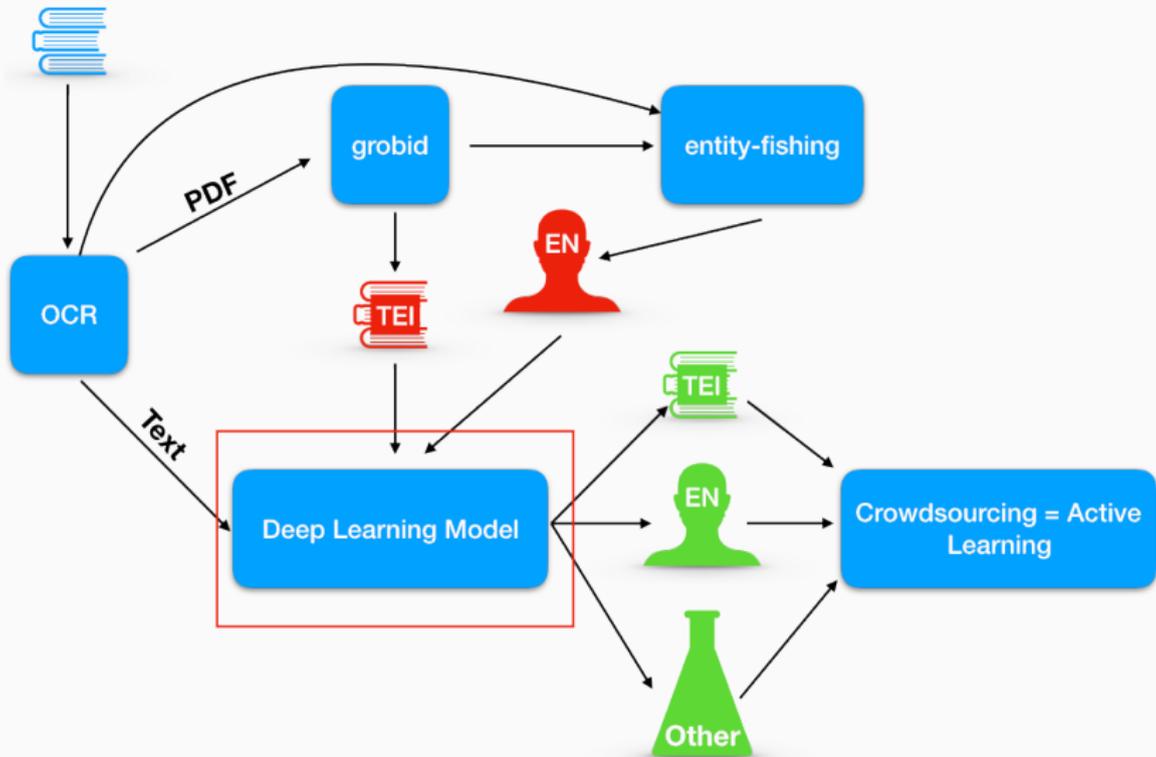
# Workflow



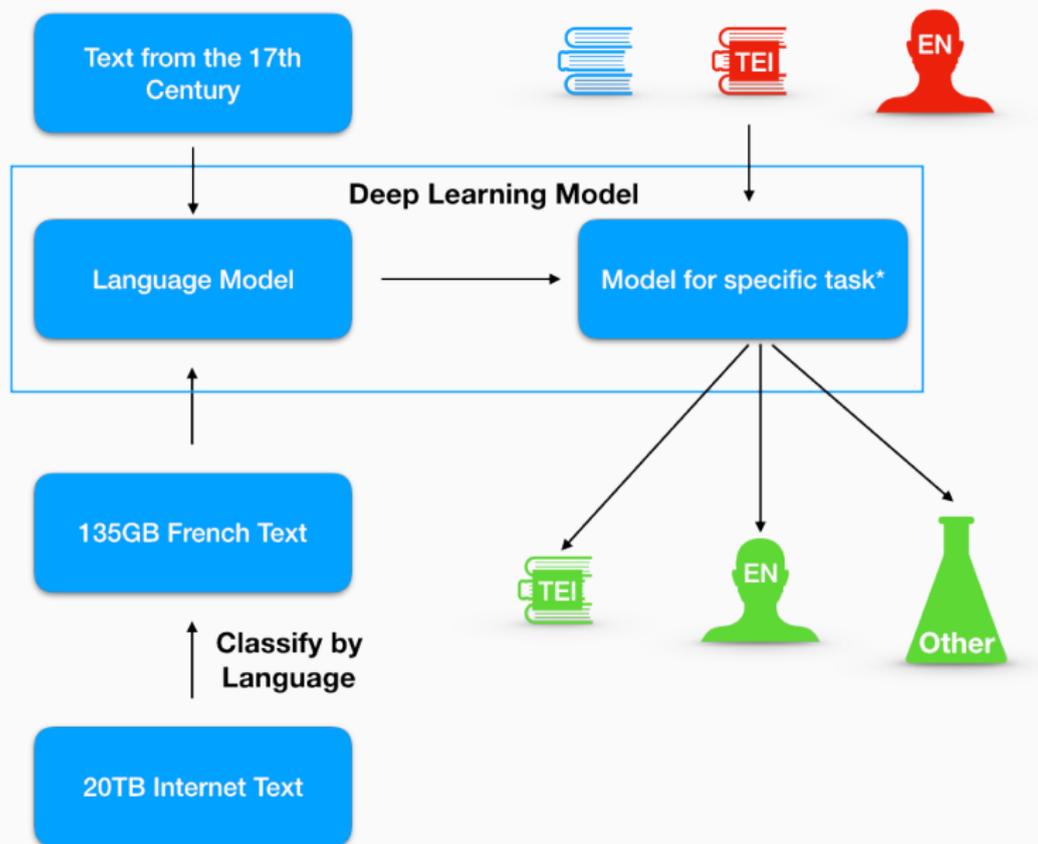
# OCR Model



# Workflow

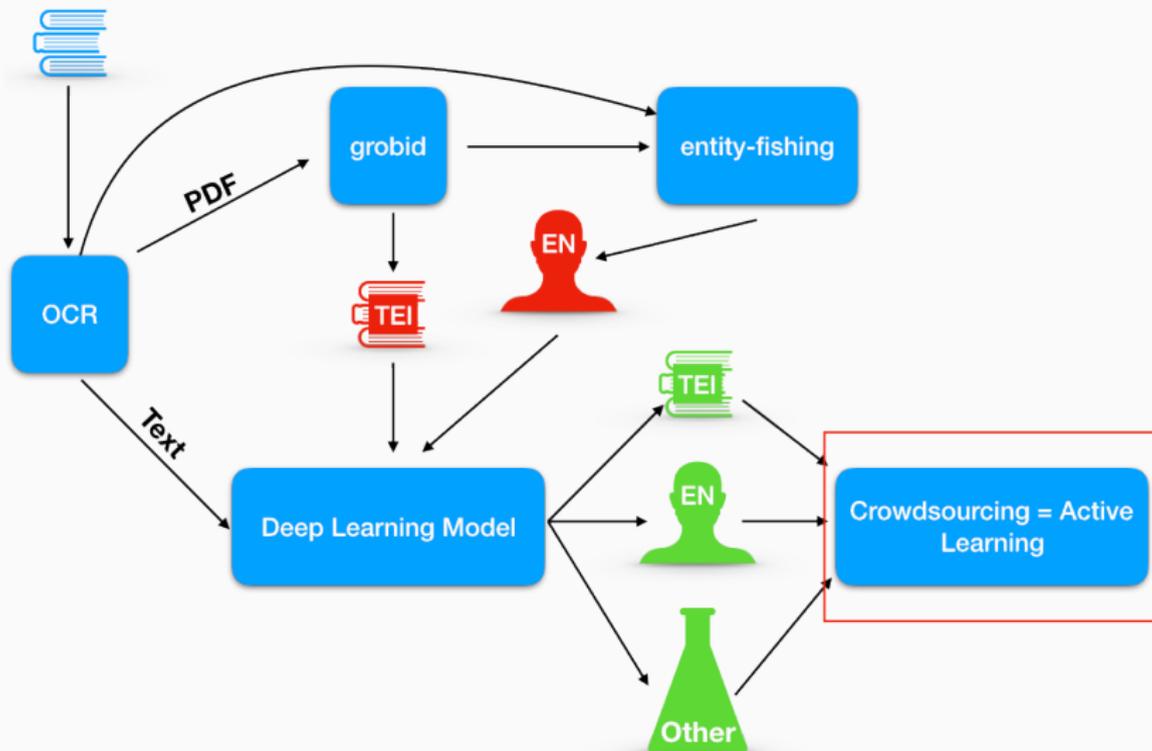


# Deep Learning Model

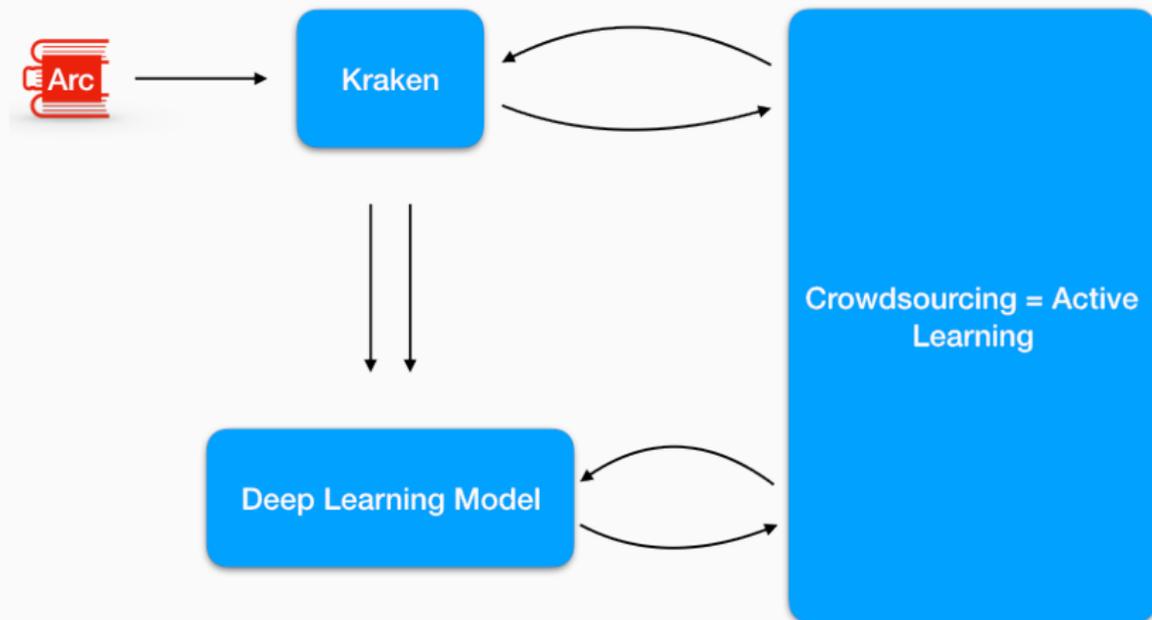


\*Changes by task

# Workflow



# Crowdsourcing Model



# Preparing the Dictionnaire Universel for Automatic Enrichment

---

# Recommendations

- **Annotate intelligently for not so intelligent models!**  
At the beginning, **only** entries for the **letter C** were transcribed.  
Our OCR engine learned that every entry starts with C.  
**Randomise your annotations!**
- **TEI is flexible**  
TEI is flexible to some extent, **agree** on how to annotate, as multiple annotations schemes reduce the precision of the model.
- **Get a lot of text in the target language [6]**  
Then get a lot of data of the target period or about the target topic.



Thank you!



H. Bots, L. Van Lieshout, and H. B. de Beauval.

*Contribution à la connaissance des réseaux d'information au début du XVIIIe siècle: Henri Basnage de Beauval et sa correspondance à propos de l'" Histoire des ouvrages des savans", 1687-1709: publication annotée de quelque cent lettres et index thématique et analytique, volume 3.*

Amsterdam, Holland University Press, 1984.



J. Considine.

*Academy Dictionaries 1600–1800.*

Cambridge University Press, 2014.



A. Furetière.

*Dictionnaire Universel, contenant généralement tous les mots françois tant vieux que modernes, & les termes des sciences et des arts.*

Arnoud et Reinier Leers, 1701.

Par Feu Messire Antoine Furetière,... 2e édition revue, corrigée et augmentée par M. Basnage de Bauval.



M. Khemakhem, L. Foppiano, and L. Romary.

**Automatic extraction of tei structures in digitized lexical resources using conditional random fields.**

*electronic lexicography, eLex 2017, Sept. 2017.*



M. Khemakhem, A. Herold, and L. Romary.  
**Enhancing usability for automatically structuring digitised dictionaries.**

*GLOBALEX workshop at LREC 2018, May 2018.*



P. J. Ortiz Suárez, B. Sagot, and L. Romary.  
**Asynchronous Pipelines for Processing Huge Corpora on Medium to Low Resource Infrastructures.**

*In 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Cardiff, United Kingdom, July 2019.*



A. Rey.  
***Antoine Furetière, imagier de la culture classique, Le dictionnaire universel d'Antoine Furetière, volume 1.***

Paris: SNL - Le Robert, 1978.